Universitat Politècnica de Catalunya

Programa de Doctorat:

Automàtica, Robòtica i Visió

Tesi Doctoral

Disseny i modelització d'un sistema de gestió multiresolució per a sèries temporals

Aleix Llusà Serra

Direcció:

Teresa Escobet Canal i Sebastià Vila-Marta

Octubre de 2015

Edició impresa: octubre de 2015.
Primera versió: 1.0.0 (composta a 9 d'octubre de 2015).

Amb el suport de la Universitat Politècnica de Catalunya (UPC).

Aleix Llusà Serra
Departament de Disseny i Programació de Sistemes Electrònics de la Universitat Politècnica de Catalunya (DiPSE–UPC)
Escola Politècnica Superior d'Enginyeria de Manresa (EPSEM), Av. de les Bases de Manresa, 61-73, 08242 Manresa (Barcelona), CATALUNYA
`aleix@dipse.upc.edu`

El codi font LaTeX del document es troba a `http://escriny.epsem.upc.edu/projects/rrb/`

Resum

Actualment és possible d'adquirir una gran quantitat de dades, principalment gràcies a la facilitat de disposar de sistemes de monitoratge amb grans xarxes de sensors. Això no obstant, no és tan senzill de gestionar posteriorment totes aquestes dades. A més, també cal tenir en compte com s'emmagatzemen aquestes dades.

D'una banda, l'adquisició de valors d'una variable al llarg del temps es formalitza com a sèrie temporal. Així, hi ha multitud d'algoritmes i metodologies d'anàlisi de sèries temporals que descriuen com extreure informació de les dades. D'altra banda, l'emmagatzematge i la gestió de les dades es formalitza com a sistemes de gestió de bases de dades (SGBD). Així, hi ha sistemes informàtics dedicats a inferir la informació que un usuari vol consultar. Aquests sistemes són descrits per models lògics formals, entre els quals el model relacional n'és la referència principal.

En aquesta tesi dissertem sobre el fet d'emmagatzemar només aquella part de les dades originals que conté una certa informació seleccionada. Aquesta selecció de la informació es duu a terme mitjançant el resum de diferents resolucions de les dades, cadascuna de les quals bàsicament són agregacions de les dades a intervals de temps periòdics. A aquesta tècnica l'anomenem *multiresolució*.

La multiresolució s'aplica a les sèries temporals. Com a resultat, s'obtenen subsèries temporals de mida finita i amb la informació resumida. Per tal de gestionar les sèries temporals, s'utilitzen SGBD específics anomenats sistemes de gestió de bases de dades per a sèries temporals (SGST). Així doncs, proposem SGST amb capacitats de multiresolució i els anomenem sistemes de gestió de bases de dades per a sèries temporals multiresolució (SGSTM). De la mateixa manera que en els SGBD, formalitzem un model pels SGST i pels SGSTM.

A causa de la naturalesa de variable capturada al llarg del temps, en l'adquisició de les sèries temporals apareixen propietats problemàtiques. Els SGSTM tenen en compte algunes d'aquestes propietats com:

- La sincronització dels rellotges en els diferents sistemes d'adquisició.
- L'aparició de dades desconegudes perquè no s'han pogut adquirir o perquè són errònies.
- La gestió d'una quantitat enorme de dades, i que a més segueix creixent al llarg del temps.
- Les consultes amb dades que no s'han recollit de manera uniforme en el temps.

Ara bé, els SGSTM són uns sistemes que emmagatzemen unes dades segons una selecció d'informació i descarten les que no es consideren importants. Per tant, prèviament a l'emmagatzematge, cal decidir els paràmetres de selecció de la informació. Per tal d'avaluar la qualitat d'aquests sistemes, depenent dels paràmetres que s'escullin, es pot utilitzar la teoria de la informació. En aquest sentit, la multiresolució es pot considerar com una tècnica de compressió amb pèrdua. Així doncs, introduïm una reflexió sobre com avaluar l'error que es comet amb la multiresolució en comparació amb disposar de totes les dades originals.

Com es diu actualment en l'àmbit dels SGBD, un mateix sistema no pot ser adequat per a tots els contextos. A més, els sistemes han de tenir en compte un bon rendiment en altres recursos a part del temps de computació, com per exemple la capacitat finita, el consum d'energia o la transmissió per la xarxa. Així doncs, dissenyem diverses implementacions del model dels SGSTM. Aquestes implementacions exploren diverses tècniques de computació: computació incremental seguint el flux de dades, computació paral·lela i computació de bases de dades relacional.

En resum, en aquesta tesi dissenyem els SGSTM i en formalitzem un model. Els SGSTM són útils per a emmagatzemar sèries temporals en sistemes amb capacitat finita i per a precomputar la multiresolució. D'aquesta manera, permeten disposar de consultes i visualitzacions immediates de les sèries temporals de forma resumida. Això no obstant, impliquen una selecció de la informació que cal decidir prèviament. En aquesta tesi proposem consideracions i reflexions sobre els límits de la multiresolució.

Abstract

Nowadays, it is possible to acquire a huge amount of data, mainly due to the fact that it is easy to build monitoring systems together with big sensor networks. However, data has to be managed accordingly, which is not so trivial. Furthermore. the storage for all this data also has to be considered.

On the one hand, time series is the formalisation for the process of acquiring values from a variable along time. There is a great deal of algorithms and methodologies for analysing time series which describe how information can be extracted from data. On the other hand, *Data Base Management System* (DBMS) are the formalisation for the systems that store and manage data. That is, these computer systems are devoted to infer the information that a given user may query. These systems are formally defined by logic models from which the relational model is the main reference.

This thesis is a dissertation on the hypothesis to store only parts of original data which contain selected information. This information selection involves summarising data with different resolutions, mainly by aggregating data at periodic time intervals. We name *multiresolution* to this technique.

Multiresolution is operated on time series. The results are time subseries that have bounded size and summaries of information. Particular DBMS are used for managing time series, then they are called *Time Series Data Base Management System* (TSMS). In this context, we define TSMS with multiresolution capabilities, which we call *Multiresolution Time Series Data Base Management System* (MTSMS). Similarly to how it is done for DBMS, we formalise a model for TSMS and for MTSMS.

The acquisition of time series presents troublesome properties owing to the fact of variable acquired along time. In MTSMS we consider some of this properties such as:

- The clock synchronisation for different acquisition systems.
- Unknown data when data has not been acquired or when it is erroneous.
- A huge amount of data to be manage. Moreover it increases as more data gets acquired.
- Queries with data that has not been acquired regularly along time.

MTSMS are defined as systems to store data by selecting information and so by discarding data that is not considered important. Therefore, the parameters for selecting information must be decided previously to storing data. The information theory is the base for measuring the quality of theses systems, which depends on the parameters chosen. Regarding this, multiresolution can be considered as a lossy compression technique. We introduce some hypothesis on measuring the error caused by multiresolution in comparison with the case of having all the original data.

Paraphrasing a current opinion in DBMS, the same system can not be adequate for all the different contexts. In addition, systems must consider performance in a variety of resources apart from computing time, such as energy consumption, storage capacity or network transmission. Concerning this, we design different implementations for the model of MTSMS. These implementations experiment with various computing methodologies: incremental computing along the data stream, parallel computing and relational databases computing.

Summarising, in this thesis we formalise a model for MTSMS. MTSMS are useful for storing time series in bounded capacity systems and in order to precompute the multiresolution. In this way they can achieve immediate queries and graphical visualisations for summarised time series. However, they imply an information selection that has to be considered previously to the storage. In this thesis we consider the limits for the multiresolution technique.

"I tenim un exèrcit, un exèrcit que fa tres-cents anys que ens defensa, i gràcies a ell no hem estat destruïts com a poble. I aquest exèrcit es diu cultura i els nostres soldats són mestres, actors, escriptors, científics, investigadors... [...] arreu dels Països Catalans. I amb aquest exèrcit farem de la nostra nació una terra lliure!"

Francesc Ribera 'Titot'

Agraïments

Vull agrair especialment als dos directors, en Sebastià i la Teresa, tota la seva ajuda per a poder arribar fins aquí, i també tota la paciència que han tingut. A més voldria estendre aquest agraïment a tot el departament de Dispositius i Programació de Sistemes Electrònics, l'Escola Politècnica Superior d'Enginyeria de Manresa i la Càtedra de Programari Lliure, en especial al Jordi Bofill.

Aquesta tesi s'ha dut a terme amb el suport de la Universitat Politècnica de Catalunya (UPC) mitjançant una beca de Formació de Personal Universitari (FPU-UPC). S'ha participat en el projectes d'investigació subvencionats pel Ministerio de Economia y Competitividad TEC2012-35571 *Nuevas aplicaciones del principio superregenerativo a comunicaciones por radiofrecuencia*, DPI2011-26243 *System Health Management and Reliable Control of Complex Systems* i DPI2014-58104-R *Control basado en la saluda y la resilencia de infraestructuras críticas y sistemas complejos.*

També en la darrera fase he rebut el suport dels companys d'Iskra. Ànims que tenim molta feina per fer.

Més abstractament, vull agrair a tots els que construeixen eines de programari lliure com GNU, LaTeX, Subversion, Python, etc. Sense aquestes eines, aquesta tesi no hauria estat possible. Particularment, s'ha estat subscrit a les llistes de debat de RRDtool i del Third Manifesto, que han servit de motivació durant l'elaboració d'aquesta tesi.

I en referència a les eines, cal no oblidar l'eina més important en aquesta tesi: la llengua. Aquesta tesi ha estat elaborada segons la Gramàtica de la Llengua Catalana provisional. Així doncs, voldria agrair a tots aquells que es dediquen a normalitzar i a normativitzar la llengua: els grans mestres de la llengua com en Pompeu Fabra, en Gabriel Bibiloni, l'Institut d'Estudis Catalans, el magnífic projecte de l'Optimot... però sobretot la tasca pacient que desenvolupa el Consorci per a la Normalització Lingüística i en especial al savi ensenyament de la Teresa Vila.

Dedicada a la Núria i al Lluc: "perquè creixi fort, sa i valent".

Índex

Part I.

Introducció

1. Introducció

Aquests darrers anys la tecnologia digital ha transformat la forma en què ens relacionem amb l'entorn i adquirim nou coneixement. Conceptes recentment apareguts com ciutat intel·ligent (*smart city*) o comptador intel·ligent (*smart meter*) són alguns dels exemples que fan visible la transformació tecnològica que vivim. Aquest nous paradigmes han incrementat el nombre de sensors actius i passius que recullen informació de l'entorn i de l'estat dels sistemes i, entre altres, han convertit els usuaris de dispositius electrònics en consumidors actius. Com a conseqüència, han aparegut grans conjunts de dades digitals per a les quals es requereixen equips informàtics que les capturin, emmagatzemin, manipulin, cerquin i visualitzin. Aquesta nova tecnologia ha quedat associada al nom de *Big Data*.

En el marc d'aquesta nova situació sorgeixen preguntes com: cal realment emmagatzemar totes les dades?, cal emmagatzemar-les indefinidament en el temps?, o, es viable tan sols emmagatzemar la informació útil i durant un temps restringit?. En aquesta tesi dissertem sobre un mètode, anomenat multiresolució, capaç de seleccionar i d'emmagatzemar la informació d'una forma determinada. Avaluem com aquesta metodologia gestiona la informació i els seus límits.

Contextualitzem la multiresolució en dos àmbits: les sèries temporals i els sistemes de gestió de bases de dades (SGBD). En aquesta tesi unim conceptes d'aquests dos àmbits i proposem solucions per a alguns dels problemes que en resulten.

Les sèries temporals Les sèries temporals es defineixen com a col·lecions d'observacions d'un mateix fenomen al llarg del temps. Les sèries temporals són útils per a formalitzar dades adquirides pels sensors, és a dir conjunts de valors mesurats en un instant de temps determinat. Aquestes sèries temporals contenen informació de l'entorn monitorat i poden ser analitzades per a predir l'evolució dels sistemes, per a detectar patrons de comportament, per a detectar anomalies, per a reconstruir els històrics, etc.

La recerca en anàlisi de sèries temporals ha augmentat en la darrera dècada, tal com explica Fu [49]. Com es mostra en el capítol 2 d'aquesta memòria, hi ha multitud de metodologies i algoritmes per a solucionar els problemes que presenten les sèries temporals. Un dels problemes de la gestió de sèries temporals és conseqüència del fet que són dades voluminoses i per tant són complicades d'emmagatzemar i de tractar [49, 114]. Aquest problema és especialment crític en el disseny de sistemes

integrats petits [135], els recursos dels quals són limitats: capacitat d'emmagatzematge, consum d'energia, temps de processament i capacitat de les comunicacions. Una altre problema ocorre quan les dades no han estat adquirides equi-espaiades en el temps, ja que alguns algoritmes d'anàlisi de sèries temporals ho requereixen.

Contextualitzem el nostre estudi de les sèries temporals en l'àmbit del monitoratge en què les variables adquirides mesuren una variable contínua en el temps, són aleatòries, el temps d'adquisició pot ser irregular, etc. En l'anàlisi de sèries temporals a vegades reben estudis més particulars, focalitzant i simplificant l'estudi en algunes propietats específiques de les dades. Per exemple a vegades les sèries temporals es redueixen a seqüències temporals en les quals només importa l'ordre en què s'han adquirit i el període d'adquisició és constant. El nostre estudi tracta les sèries temporals des del punt de vista genèric temporal en què, a més, cal saber la posició de temps absoluta que ocupa cada dada i la distància de temps entre els valors.

Els SGBD Els SGBD són els sistemes informàtics encarregats d'emmagatzemar i gestionar dades de forma genèrica, els quals s'inscriuen en l'àmbit dels sistemes d'informació. Els conceptes dels SGBD es basen en models matemàtics formals que defineixen l'estructura dels objectes i les operacions que els usuaris perceben de forma abstracta. El model formal de referència dels SGBD és el model relacional [23].

Els SGBD habituals fins a l'actualitat han estat els que disposen d'un llenguatge de consultes anomenat *Structured Query Language* (SQL), tot i que en els darrers anys han aparegut altres sistemes, anomenats *NoSQL* i *NewSQL*, per tal de millorar el rendiment i la flexibilitat dels SQL [5, 117, 120, 139]. Alguns autors [42, 109, 120, 139] comenten els aspectes en què cal millorar els SGBD per tal de tractar les sèries temporals i fan èmfasi en el fet que és necessari un model que enllaci el coneixement dels SGBD amb el de l'anàlisi de sèries temporals. Sobretot, l'adquisició contínua de nous valors és un repte a l'hora d'emmagatzemar i analitzar les sèries temporals [67].

Un altre repte dels SGBD és que gestionin adequadament i amb coherència l'atribut temporal de les sèries temporals. En els SGBD també es gestionen històrics temporals, és a dir l'evolució de les dades al llarg del temps. Aquest és un problema similar a les sèries temporals pel que fa a que tots dos treballen amb atributs temporals, tot i així pertanyen a dues categories diferents de dades i no poden ser tractats de la mateixa manera [3, 109]. Ara bé, els històrics temporals han estat llargament estudiats en els SGBD i finalment s'han formalitzat com a intervals temporals dins del model relacional [33]. Les sèries temporals necessiten consolidar un model similar en què se n'estudiïn les propietats i els requisits específics [42, 110].

En aquest treball estudiem els SGBD que treballen específicament amb sèries temporals, aleshores els anomenem sistemes de gestió de bases de dades per a sèries

temporals (SGST). De la mateixa manera, dissenyem SGST amb capacitats de multiresolució, els quals anomenem sistemes de gestió de bases de dades per a sèries temporals multiresolució (SGSTM). Formulem el model dels SGSTM per la qual cosa necessitem també formular el model dels SGST.

La multiresolució La multiresolució resumeix la informació d'una sèrie temporal mitjançant un conjunt de resolucions. Cada resolució correspon a un atribut i a un període de temps de la sèrie temporal. El concepte de multiresolució prové d'un estudi profund d'un sistema anomenat RRDtool [97]. El nostre objectiu és descriure els conceptes principals de la multiresolució de manera abstracta per a obtenir un model formal dels SGSTM.

La multiresolució implica una selecció d'informació i, per tant, alhora implica una pèrdua d'informació. Com a conseqüència, l'usuari ha de determinar un esquema de multiresolució per a cada sèrie temporal que vulgui gestionar amb un SGSTM. A partir del model dels SGSTM reflexionarem sobre com gestionen la informació i sobre quins efectes tenen diferents esquemes de multiresolució.

El model de SGSTM es dissenya de forma genèrica per a permetre'n implementacions vàries. Així, hi ha vàries aproximacions possibles per a computar la multiresolució: limitar l'emmagatzematge pensant en sistemes petits, precomputar el resultat per a disposar-ne visualitzacions immediates, acumular totes les dades i processar-les en temps diferit aprofitant computació paral·lela, repartir el temps de computació durant el mateix procés d'adquisició, etc. En aquest document explorarem aquestes possibilitats dels SGSTM.

1.1. Contribució d'aquesta tesi

A continuació resumim les principals contribucions d'aquesta tesi:

- Un SGSTM emmagatzema les sèries temporals de forma comprimida. És un resum de l'evolució dels atributs de la sèrie temporal al llarg del temps i amb diferents períodes temporals. És una solució de compressió amb pèrdua i per tant l'esquema de multiresolució s'ha de decidir prèviament per a cada context.
- Per a resumir cada atribut de la sèrie temporal s'utilitzen funcions d'agregació i funcions de representació, els quals es formalitzen com a objectes independents en el model. D'aquesta manera, els usuaris poden definir diferents operadors considerant la semàntica de les sèries temporals en cada context diferent. Per exemple, els atributs es poden calcular mitjançant funcions d'agregació d'estadístics com la mitjana o el màxim.

- El model opera coherentment amb la dimensió temporal de les sèries temporals. A més, té en compte les irregularitats del mostreig, el tractament i validació de dades i diverses interpretacions de les sèries temporals
- El model es basa fermament en l'àlgebra de conjunts i la del model relacional com a teoria formal dels sistemes d'informació.
- Es tenen en compte les diverses possibilitats de computació de la multiresolució i es desenvolupen diverses implementacions del model. Es desenvolupa una implementació de referència, una per a computació paral·lela i una amb llenguatge acadèmic relacional.
- S'introdueix el problema d'avaluar la qualitat de la multiresolució. És a dir el problema de determinar quina selecció i quina pèrdua d'informació hi ha quan s'aplica la multiresolució a una sèrie temporal i com realitzar consultes aproximades a la informació original.

Aquesta tesi ha derivat en altres publicacions, que detallem a continuació:

- Report de recerca [85] conjuntament amb la doctora Teresa Escobet Canal i el doctor Sebastià Vila Marta, publicat el 17 de desembre de 2012 al departament de Disseny i Programació de Sistemes Electrònics de la Universitat Politècnica de Catalunya. És un report on consten les mancances que tenen els SGBD per a les sèries temporals, les propietats i requisits que haurien de complir i la idea bàsica de la proposta d'un nou model multiresolució per a sèries temporals.
- Ponència a congrés [84] conjuntament amb la doctora Teresa Escobet Canal i el doctor Sebastià Vila Marta, realitzada a l'*International Conference on Artificial Intelligence, Knowledge Engineering and Data Bases* (AIKED '13) a Cambridge, UK, els dies 20–22 de febrer de 2013. Es dóna a conèixer de forma resumida el model multiresolució que dissenyem.
- Article a *Information Systems* [86] presentat conjuntament amb la doctora Teresa Escobet Canal i el doctor Sebastià Vila Marta. Un cop s'ha completat el disseny del model de SGSTM, s'ha escrit compactament en format article per a l'àmbit de les bases de dades. En aquest article també s'ha inclòs el disseny de la implementació de referència dels model i la motivació del treball comparada amb recerques similars. Aquest article es publicarà a la revista *Information Systems* del març de 2016 tot i que ja es troba disponible en línia a la mateixa revista.
- Les implementacions són de programari lliure i es poden trobar a [83, `http://escriny.epsem.upc.edu/projects/rrb/repository/show/src`]. Tot i que són experimentals i tenen nivell acadèmic, mostren el funcionament correcte del model.

1.2. Estructura del document

Aquest document és el resultat de la recerca en un model de multiresolució per a les sèries temporals. S'estructura en cinc parts principals.

En una primera part, que inclou aquest capítol 1 d'introducció, es presenta el context i els objectius de la recerca. En el capítol 2 es descriu l'estat actual de la recerca i treballs similars.

En una segona part es dissenya i es formula el model. En el capítol 3 s'introdueixen els conceptes de model de SGBD i l'abast de les sèries temporals i la multiresolució que utilitzem. El model es dissenya en dues parts: el model dels SGST en el capítol 4 i el model dels SGSTM en el capítol 5.

En una tercera part s'exploren variacions i reflexions sobre el model presentat. En el capítol 6 s'introdueixen les consideracions que es faran sobre el model i s'exploren petites variacions en l'emmagatzematge de les sèries temporals. En el capítol 7 es formula la multiresolució com una funció que a partir d'una sèrie temporal resulta en una nova sèrie temporal o conjunts de sèries temporals comprimides. En el capítol 8 es dissenyen sistemes que utilitzin alhora SGST i SGSTM. En el capítol 9 es reflexiona sobre el problema de la qualitat de la multiresolució, és a dir s'avalua quina selecció d'informació fa un determinat esquema de multiresolució.

En una quarta part s'experimenta i es dissenyen les implementacions del model. En el capítol 10 s'introdueixen les diferents implementacions i es comenten les particularitats de cada una. En el capítol 11 es dissenya la implementació de referència amb llenguatge Python. En el capítol 12 es dissenya una implementació amb computació distribuïda i paral·lela amb la tècnica MapReduce i en el sistema Hadoop. En el capítol 13 es dissenya una implementació amb el llenguatge acadèmic relacional Tutorial D. En el capítol 14 s'exemplifica l'ús de les implementacions amb dades reals.

En una cinquena part es conclou el document. Particularment en el capítol 15 es resumeix la dissertació i s'expressen les conclusions que se'n poden treure, i en el capítol 16 es proposen altres investigacions que serien interessants a partir d'aquesta recerca.

Finalment, com a materials de referència es recull la bibliografia citada, les abreviacions i la nomenclatura utilitzades i s'ofereixen índexs per a les entitats remarcables –figures, taules, llistats, definicions i exemples– les quals numerem precedides del número de secció per a facilitar-ne la localització.

2. Estat de la qüestió

En aquest capítol resumim la base teòrica per als models que proposem i l'avantguarda de l'estat de la qüestió. S'estructura en tres parts:

- Sèries temporals. L'anàlisi, l'adquisició, el monitoratge, l'emmagatzematge i la gestió de sèries temporals.
- Sistemes de gestió de bases de dades (SGBD). El model relacional i sistemes actuals.
- Sistemes i projectes similars. Resum d'altres sistemes que gestionen sèries temporals.

2.1. Sèries temporals

Una sèrie temporal és un conjunt de valors cadascun dels quals té associat un instant de temps diferent. Tradicionalment s'anomenen sèries temporals tot i que també s'accepta la denominació de seqüències temporals [57].

Les sèries temporals s'emmarquen dins l'àmbit més genèric del que es coneix com a *dades temporals*. Les dades temporals són col·leccions de dades arbitràries que estan associades a la dimensió temps. Dins del concepte de dades temporals s'hi encabeixen col·leccions de dades de diversa natura. En funció de com un valor queda vinculat amb el temps, es poden diferenciar dues categories [3]:

1. La primera la formen les sèries temporals tal i com s'han definit prèviament, en la qual la dada està associada a un instant de temps.
2. La segona, que s'anomena dades bitemporals (*bitemporal data*), la formen col·leccions de dades en què cada element té dos atributs temporals: el rang de validesa, que indica l'interval de temps en que la dada és vàlida, i el temps de transacció, que indica quan es va desar la dada a la base de dades.

Aquestes dues categories de dades temporals, tot i tenir aspectes en comú, no poden ser tractades amb les mateixes eines, [109].

Les sèries temporals s'utilitzen en camps molt diversos i amb objectius molt diferents. L'ús generalitzat és per a l'anàlisi i la comprensió del comportament temporal de variables. L'evolució d'una sèrie temporal es pot representar amb un model.

Aquests models, en l'àmbit de l'enginyeria, permeten realitzar tasques relacionades amb validació de dades, diagnòstic i prognosis. Per exemple, trobem aplicacions de sèries temporals en el camp de l'avaluació de la degradació de components [137], anàlisi de l'estat dels sensors d'un vaixell [100], validació i reconstrucció de dades en xarxes de distribució d'aigua [107], classificació de valors econòmics [44], optimització de la planificació semafòrica [77], estimació del temps de viatge en autopistes [115] o transmissió d'informació en xarxes de sensors [61, 135].

L'objectiu d'aquest apartat és mostrar l'estat de l'art dels principals processos vinculats en el treball amb sèries temporals. A tal efecte s'organitza en tres subapartats.

El primer subapartat tracta de l'anàlisi de sèries temporals, que és la formalització de les tècniques que s'utilitzen per extreure informació. A vegades aquesta extracció també es coneix com descobriment de coneixement i es pot emmarcar dins de l'àrea de l'intel·ligència artificial.

El segon subapartat se centra en l'adquisició de dades. El primer requeriment d'una sèrie temporal és l'adquisició de dades. Els sistemes de monitoratge s'encarreguen de recollir dades dels sensors, periòdicament o en base a esdeveniments. Els problemes que es donen durant l'adquisició generen defectes específics en les sèries temporals que cal analitzar i tractar convenientment.

El tercer subapartat es dedica als sistemes d'emmagatzematge de sèries temporals. L'emmagatzematge de les dades i la implementació de les tècniques d'anàlisi ocorre en els SGBD. Aquests s'encarreguen de l'organització correcte de la informació i de respondre a les operacions de consulta. Les sèries temporals necessiten un tractament específic per part d'aquests sistemes.

2.1.1. Anàlisi de sèries temporals

L'anàlisi de sèries temporals consisteix en l'aplicació de metodologies i d'algoritmes que permeten tasques com per exemple l'extracció de característiques o obtenció de models. Aquestes tècniques es recullen en el que es coneix amb el nom de mineria de sèries temporals (*time series data mining*). La mineria de dades, en la qual s'inscriuen les sèries temporals, és l'estudi d'algoritmes específics per a extreure patrons de comportament de les dades i és una etapa del procés general de descobriment de coneixement a les bases de dades (*knowledge discovery in databases*) [47, 79]

Actualment, les sèries temporals es consideren com un dels deu problemes prioritaris en la mineria de dades [134]. Tal com esmenta Fu [49] en un article recent, la recerca en mineria de sèries temporals s'ha incrementat en la darrera dècada. L'objectiu principal és reduir la mida de les sèries temporals per tal de disminuir el temps de processat de les dades. Fu resumeix l'estat actual de la mineria de sèries temporals de forma exhaustiva i conclou que encara queden molts problemes per investigar i

resoldre. La recerca en tasques de mineria ha estat intensa però es necessita millorar la representació de sèries temporals, ja que és fonamental per a reduir la mida de les dades.

Segons Keogh i Kasetty [64], les quatre tasques que centren l'atenció de la recerca actual de sèries temporals són l'indexat (*indexing*), que treballa amb una estructura comprimida de les dades; l'agrupament (*clustering*), que agrupa les dades segons la similitud entre elles per tal de descobrir patrons; la classificació (*classification*), que etiqueta les dades segons les característiques que presentin; i la segmentació (*segmentation*), que parteix una sèrie temporal en subseqüències. A més, Keogh i Kasetty comparen alguns algoritmes experimentals i recomanen a la comunitat de mineria de sèries temporals que segueixi el seu estudi com a punt de referència per avaluar el rendiment d'algoritmes similars.

Un pas comú previ a les quatre tasques anteriors és el de representació de la sèrie temporal. Les sèries temporals són discretes, són valors en punts de temps discrets, i la representació és el model de funció que aproxima la sèrie temporal a la seva naturalesa contínua original. La mineria de sèries temporals aprofita la representació per reduir la mida de les sèries temporals. Keogh et al. [70], cita vàries representacions per les sèries temporals com per exemple *Fourier Transforms*, *Wavelets*, *Symbolic Mappings* o *Piecewise Linear Representation* (PLR), però assenyala aquesta última com la representació més utilitzada. La PLR [66, 67] és una representació definida a trossos lineals que s'aproxima a la sèrie temporal. Els trossos podrien ser polinomis de qualsevol grau, però la manera més comuna de representar sèries temporals és amb funcions lineals ja que és més propera a la visió de l'ésser humà que segmenta les corbes en línies rectes. A banda de la PLR, Keogh et al. [68, 69] exploren altres representacions de sèries temporals per tal de reduir la mida d'una sèrie temporal i poder-la indexar més fàcilment. Proposen dues tècniques eficients en el càlcul: la *Piecewise Aggregate Aproximation* i la *Adaptive Piecewise Constant Approximation*, ambdues basades en la representació a trossos constants de la sèrie temporal. D'aquestes dues tècniques, Keogh et al., Keogh et al. conclouen que mantenen una bona aproximació a la sèrie temporal i que a més tenen molt menys cost de càlcul que altres de més complicades, com ara la *Discrete Fourier Transform*, la *Singular Value Decomposition* o la *Discrete Wavelet Transform*.

Altres representacions també aproximen una sèrie temporal a trossos però generalitzen la funció d'aproximació. Per exemple Last et al. [79] proposa aproximar una sèrie temporal partint-la en subintervals i calculant per a cada un la funció que més s'hi aproxima. Un altre àmbit on s'aplica l'anàlisi de sèries temporals és en teoria del senyal. Per tant també és possible utilitzar les representacions habituals en teoria del senyal, per exemple es pot representar una sèrie temporal amb una funció graó (*step* o *staircase function*); és a dir, amb una funció definida a trossos constant (*piecewise constant representation*).

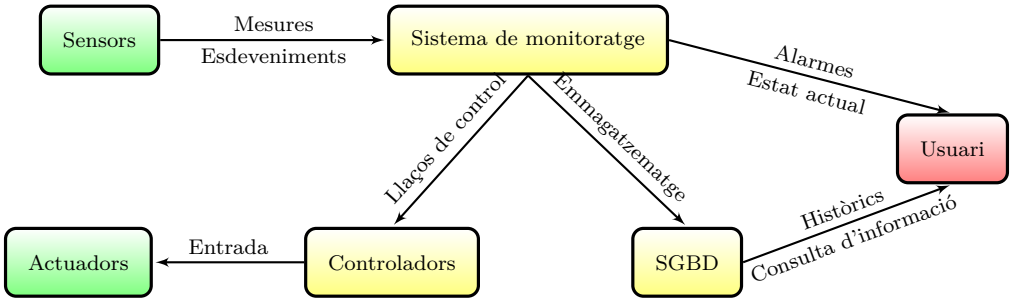


Figura 2.1.: SCADA: de l'adquisició de dades fins a informar l'usuari

2.1.2. Adquisició i monitoratge de sèries temporals

Els sistemes de monitoratge són una part important d'interacció entre un procés i els usuaris, entenent com a procés qualsevol sistema físic, químic, ambiental, etc. del qual es pugui recollir informació continuada, ja sigui de forma periòdica o en funció d'esdeveniments. Principalment, aquests sistemes s'encarreguen de recollir dades, conèixer l'estat actual del procés i informar a l'usuari. Els sistemes de monitoratge constitueixen la part principal dels sistemes de supervisió, control i adquisició (SCADA, de l'angl. *Supervisory Control And Data Acquisition*). Un SCADA és un sistema encarregat de recollir i centralitzar les dades de manera periòdica en el temps.

Com exemple, a la figura 2.1, es mostren els blocs principals d'un SCADA. El monitor adquireix dades dels sensors. Les dades poden ser valors de mesures o estats del procés adquirits com a esdeveniments. Fent referència a la classificació de dades temporals de Aßfalg [3], en general les mesures es poden entendre com a sèries temporals i els esdeveniments com a dades bitemporals.

En el cas de sistemes controlats o automatitzats, les dades adquirides poden ser utilitzades per comandar o modificar el funcionament del procés. Aleshores s'incideix en diferents nivells des de llaços de control modificant directament un accionament, fet que no sol ser habitual ja que els llaços de control solen realitzar-se els sistemes electrònics que resideixen prop dels sistemes controlats, fins a gestió de modes de funcionament i coordinació entre màquines.

L'ús generalitzat dels sistemes de monitoratge és el de proporcionar informació de l'estat actual del procés. També disposen de la possibilitat de generar alarmes senzilles com per exemple que no s'han pogut adquirir les dades o que el sensor ha assolit un valor crític. Per a usuari ens referim tant a un usuari humà com a un altre sistema supervisor dotat amb intel·ligència artificial. Per a càlculs més complicats amb les dades, els sistemes de monitoratge utilitzen SGBD. Mitjançant els SGBD, s'emmagatzemen les dades en bases de dades i posteriorment l'usuari les

consulta per observar els històrics o per obtenir informació i elaborar coneixement a partir de les dades emmagatzemades.

La figura 2.1 presenta una visió centralitzada de l'adquisició de dades. Ara bé, els sistemes de monitoratge internament poden tenir estructura distribuïda quan els sensors tenen suficient capacitat de processament, com per exemple les xarxes de sensors. En aquests casos els monitors distribueixen parts al sensors, sobretot pel que fa als SGBD que passen a tenir un paper més rellevant en la comunicació.

Un dels camps recents on l'adquisició de sèries temporals hi juga un paper fonamental és el de les xarxes de sensors. L'abaratiment del maquinari permet monitorar el procés amb grans quantitats de sensors intel·ligents [61, 135], els quals tenen processador i sistemes de comunicació incorporats però tenen recursos limitats pel que fa a transmissió, energia i processament i estan sotmesos a la incertesa dels sensors. Així doncs, el problema de les xarxes de sensors rau en estudiar l'ús eficient d'aquests recursos, per la qual cosa actualment trobem dues propostes. Una solució consisteix en transmetre la informació a un node central comprimint-la tant amb agregacions o estadístics com amb aproximacions [35]. Una altra solució consisteix en tenir les dades distribuïdes en diferents sensors i decidir com s'ha de resoldre cada consulta tenint en compte que el processament local és més barat que la comunicació [9, 53, 73, 135].

Problemes en el monitoratge

Els sistemes de monitoratge habitualment presenten problemes derivats de la recol·lecció de dades. A la bibliografia dels sistemes de monitoratge es remarquen bàsicament tres problemes.

1. El primer problema és la gestió d'una quantitat enorme de dades.

   Un sistema de monitoratge recull una gran quantitat de dades. Ara bé, l'usuari només en pot observar una petita part d'aquestes dades sincronitzat (*online*) amb el procés i les dades emmagatzemades esdevenen massa grans per a ser processades posteriorment [67]. No obstant, totes les dades han de ser analitzades ja que contenen informació interessant per a les aplicacions de les sèries temporals descrites a l'apartat anterior. S'observa que en el context de monitoratge les dades recollides es poden considerar com a sèries temporals ja que abstractament són una col·lecció de mesures.

2. El segon problema és el de la necessitat de validar les dades, és a dir comprovar que les dades siguin correctes i en cas contrari rebutjar-les o reconstruir-les.

   Quevedo et al. [107] mostren la quantitat d'informació que hi ha en els sistemes complexos de telecontrol. Aquesta informació s'obté de diversos sensors distribuïts pel camp de mesura. En el moment de recol·lecció de dades apareixen dos problemes: valors que en un instant de temps prefixat no s'han

pogut recollir i valors que són incorrectes. En el procés de gestió de dades no es poden emmagatzemar les dades amb aquests dos tipus de problema ja que aleshores els registres històrics serien inconsistents. Així doncs, cal comprovar que les dades emmagatzemades són correctes, mitjançant un procés de validació, i modificar-les en el cas que siguin incorrectes, mitjançant un procés de reconstrucció que estimi els valors correctes. Per exemple, Quevedo et al. apliquen aquests processos de validació i reconstrucció a xarxes de distribució d'aigua.

3. El tercer problema es dóna quan el període de mostreig no és regular, és a dir que les dades no es recullen de manera uniforme en el temps, però les aplicacions no ho contemplen o volen treballar amb dades a intervals regulars, també anomenat dades equi-espaiades.

   Una causa de la irregularitat es deu a que els sistemes de monitoratge informàtics sovint no són capaços de complir amb exactitud el temps de mesura sinó que presenten una certa variació, ja sigui deguda a retards en els sensors, les comunicacions o la concurrència del monitoratge amb altres tasques del sistema operatiu. Aquesta causa, però, es pot atenuar si els sensors envien el temps de mesura juntament amb el valor mesurat. Aleshores, el problema recau en la sincronització dels rellotges dels sensors que descriu Kopetz [74, cap. 3].

   En els sistemes controlats o automatitzats, el sistema de monitoratge ha d'obeir a les restriccions de temps imposades pels llaços de control. Aquestes restriccions són especialment crítiques en els sistemes de control en temps real ja que, aleshores, el sistema de monitoratge no pot imposar restriccions de temps diferents de les que s'han calculat per als llaços de control. Lozoya, Velasco i Martí [80] mostren que s'ha de vigilar amb les entrades i sortides de les tasques periòdiques als sistemes en temps real. L'actuació dels sistemes de control es degrada quan no es té en compte que les operacions d'entrada i sortida estan subjectes a fluctuacions degudes al mostreig i a latències. Aquest problema afecta als sistemes de monitoratge en dues vessants. Per una banda, els sistemes de monitoratge tenen una part de l'adquisició controlada per les aplicacions de control en temps real i per tant el període de mostreig resultant que veu el monitor no és regular. Per altra banda, les aplicacions que analitzen les dades obtingudes del monitoratge poden veure com la seva actuació es degrada si no consideren que l'adquisició de dades és irregular. Això és similar a la regressió que s'observa [80] quan en el disseny d'un controlador discret es considera que es mostreja i s'actua periòdicament però hi ha un sistema en temps real que fa fluctuar la periodicitat, sobretot si el control basa el mostreig segons els esdeveniments que ocorren.

En conclusió, per tal de gestionar la complexitat derivada de la recollida de dades i també la complexitat de les consultes posteriors per part de l'usuari, els sistemes

de monitoratge es recolzen en SGBD per gestionar l'emmagatzematge de les dades i la recuperació d'informació.

2.1.3. Emmagatzematge i gestió de sèries temporals

Els SGBD són els sistemes informàtics que s'encarreguen d'emmagatzemar informació i de permetre a l'usuari consultar-la. Més endavant, a la secció 2.2, descrivim com es formalitzen els SGBD, en aquest apartat ens centrem en les necessitats que tenen les sèries temporals dels SGBD.

Les sèries temporals es diferencien d'altres tipus de dades en el fet que els seus valors són dependents d'una variable: el temps. Com a conseqüència, qualsevol SGBD que les vulgui tractar no ho pot fer de manera independent pels valors i pel temps; ha de conservar la coherència temporal. Per poder aplicar les tècniques d'anàlisi de sèries temporals de manera eficient cal disposar de SGBD específics. Durant l'última dècada, el maquinari informàtic ha millorat tant des del punt de vista tecnològic com econòmic [35], la qual cosa ha facilitat el procés d'adquisició de dades i alhora ha ampliat la capacitat per emmagatzemar les dades. Així doncs, el volum de dades a tractar en els SGBD cada cop esdevé més crític, cosa que actualment es coneix amb el nom de *Big Data* [60].

En l'àmbit d'aplicació dels SGBD, el problema de grans quantitats de dades també es troba en altres camps, com assenyalen Mylopoulos et al. [93] quant a la necessitat de grans bases de dades de coneixement. Els SGBD que tracten aquestes dades s'anomenen *very large databases* (VLDB), els quals han de construir, accedir i gestionar la quantitat de dades de manera eficient. Ogras i Ferhatosmanoglu [99] consideren que les aproximacions que fan les VLDB estan pensades per a bases de dades estàtiques i en canvi observen que les sèries temporal normalment són dinàmiques, és a dir de naturalesa contínua i de mida no fitada. Conseqüentment, conclouen que les solucions tradicionals, les quals analitzen a posteriori i sense tenir en compte l'ordre, no es poden aplicar a les sèries temporals a causa de l'arribada seqüencial i contínua de les dades. Com a solució proposen resumir dinàmicament les sèries temporals mitjançant tècniques de compressió que s'utilitzen en altres aplicacions on hi ha bases de dades grans.

Les sèries temporals es poden emmagatzemar i gestionar en els SGBD habituals per a altres dades, com els sistemes amb SQL. Això no obstant, alguns autors [42, 109, 120, 139] consideren problemàtic l'ús de sistemes SQL com a suport per a les sèries temporals. Per tal d'incrementar-ne el rendiment i la flexibilitat, s'estan desenvolupant productes *NoSQL* o *NewSQL* [5, 117, 120, 139], encara que la naturalesa d'adquisició contínua de les sèries temporals és un repte per a emmagatzemar i analitzar en temps diferit totes les dades capturades [67].

Dreyer, Dittrich i Schmidt [42] proposen desenvolupar SGBD que implementin operacions específiques per les sèries temporals, aleshores els anomenen SGST. Consideren

que els altres SGBD no són adequats per a tractar sèries temporals, tot i que després de comparar els SGBD per a dades bitemporals i els SGST [109] troben que hi ha aspectes comuns entre tots dos sistemes. Els SGST estan optimitzats per gestionar les dades segons les operacions de temps i rotació, les quals són molt comunes en la gestió de les sèries temporals. A més també cal controlar el creixement de la base de dades i la consulta ha de ser flexible i d'alta velocitat [12]. Les propietats d'un model de SGST han estat estudiades per Segev i Shoshani [110] en la forma d'un model genèric per dades temporals. No obstant això, fins on coneixem, la recerca posterior s'ha concentrat en tasques de mineria de dades. Per exemple Last et al. [79] estudien una metodologia general per descobrir coneixement en els SGST, tant pel que fa a patrons temporals com a regles temporals, i breument noten l'existència de la proposta de Dreyer, Dittrich i Schmidt [42] pels SGST.

Altres estudis proposen tractar les sèries temporals com a tipus que tenen ordre, per exemple seqüències o matrius.

Seshadri [112] proposa que les sèries temporals són un subconjunt de les seqüències i per tant el model i les operacions per les seqüències [113] serveixen per les sèries temporals. Bonnet, Gehrke i Seshadri [9] utilitzen el model de seqüències en SGBD distribuïts per xarxes de sensors, aleshores l'estratègia de comunicació inclou agregacions de les sèries temporals en els sensors [36]. També es relaciona el model de seqüències de les sèries temporals amb els *data streams* [6, 59, 99]. Els *data streams* són dades que arriben contínuament i amb ordre temporal i es modelen com una seqüència on només s'hi poden afegir elements. Aleshores les consultes poden ser contínues, és a dir cada cop que arriba una dada nova s'actualitza incrementalment la informació. Per les sèries temporals s'utilitza en el càlcul de correlacions i prediccions de forma incremental [136] i en la cerca de patrons [7].

En els SGBD per matrius (*arrays*) destaquen els anomenats sistemes de gestió de bases de dades científiques, camp en el qual les sèries temporals hi tenen un paper de primer ordre [110, 139]. Stonebraker et al. [120] estudien les necessitats d'aquests sistemes sobretot en l'àmbit de la ciència. Kersten et al. [72] proposen un sistema molt semblant però a més integren el seu llenguatge, anomenat SciQL (SQL *for science applications*), amb la sintaxi de SQL. Zhang et al. [139] exemplifiquen detalladament l'ús de SciQL en les sèries temporals per a algunes de les seves propietats: regularitat, interpolació i cerca de correlacions.

2.2. Sistemes de gestió de bases de dades

Segons Date [22], "una base de dades és un contenidor informàtic persistent per a una col·lecció de dades". El sistemes informàtics que tracten amb bases de dades s'anomenen sistemes de gestió de bases de dades (SGBD) i tenen els objectius d'emmagatzemar informació i permetre consultar i modificar aquesta informació. Per complir aquests objectius, els SGBD ofereixen a l'usuari diferents operacions

com per exemple crear una base de dades, afegir dades o operar amb les dades emmagatzemades.

Els àmbits d'aplicació dels SGBD són varis: operacions repetitives i rutinàries de producció, anomenades *online transaction processing*; sistemes per a prendre decisions empresarials, a vegades anomenats *data warehouse*; processament de dades científiques; etc. Alguns dels avantatges de gestionar aquestes dades en bases de dades són: evitar la disgregació de la informació i tenir-la perfectament organitzada, poder compartir la mateixa informació entre diverses aplicacions, garantir la consistència i la integritat de les dades i evitar redundàncies innecessàries, afegir seguretat a la gestió de les dades o optimitzar les consultes que l'usuari sol·licita.

Els SGBD es poden descriure mitjançant teories matemàtiques que reben el nom de *model de dades*. Segons Date, "un model de dades és una definició abstracta, auto continguda i lògica dels objectes, de les operacions i de la resta que conjuntament constitueixen la màquina abstracta amb què els usuaris interaccionen. Els objectes permeten modelar l'estructura de les dades. Les operacions permeten modelar el comportament". Ara bé, Date avisa que el concepte *model de dades* també s'usa per a definir una estructura o esquema persistent de dades concreta i, per tant, cal distingir adequadament entre tots dos significats. Tal com fa Date, en aquest document parlarem de model de dades, o simplement de model, en el primer sentit de màquina abstracta. També distingeix entre els conceptes de *dades* –allò que està emmagatzemat a la base de dades– i *informació* –el significat que algú dóna a aquestes dades.

Un model de SGBD que ha s'ha consolidat i ha esdevingut un referent és el model relacional (*relational model*). L'èxit d'aquest model és degut principalment que es fonamenta en teories matemàtiques consolidades: la lògica de predicats i la teoria de conjunts [22]. En base al model relacional es va definir el llenguatge *Structured Query Language* (SQL) per operar amb bases de dades que ha esdevingut un estàndard en molts SGBD.

Els SGBD se solen dissenyar amb una arquitectura de tres nivells: el físic, el lògic i el d'usuari [22].

- El nivell d'usuari o extern agrupa les eines que tenen disponibles els usuaris per a interactuar amb la base de dades, per exemple en el cas relacional SQL pertany a aquest nivell.
- El nivell lògic o conceptual és l'abstracció formal dels conceptes dels SGBD. En aquest nivell hi pertanyen els models de dades, per exemple el mateix model relacional en el cas relacional.
- El nivell físic o intern agrupa la programació informàtica de com s'han d'emmagatzemar físicament les base dades i de com s'han d'executar les operacions. Per exemple en aquest nivell apareixen els registres de memòria, punters, mètodes d'accés als fitxers, etc., conceptes que no tenen res de relacional.

Una bona diferenciació entre els tres nivells d'arquitectura aporta independència a les dades (*data independence*) [26]. Date considera que és una de les propietats més importants que han de complir els SGBD. De forma resumida, la independència a les dades significa que el nivell lògic no ha de contenir detalls d'implementació ni parlar d'objectius de rendiment sinó que aquests són part del nivell físic. D'aquesta forma és possible canviar el nivell físic sense afectar el nivell lògic. Així doncs, un model de dades concret pot tenir diverses implementacions en el nivell físic, per exemple *PostgreSQL* [56] per al model relacional. Pascal [101] detalla algunes confusions actuals sobre la independència entre el model i la implementació.

2.2.1. Sistemes relacionals

El model relacional va ser proposat per Codd [16, 17] com una teoria abstracta de dades per tal de formalitzar els SGBD amb teories matemàtiques consolidades. El model relacional va significar un gran canvi en la recerca en SGBD ja que, a diferència dels models antics, possibilitava l'estudi dels problemes amb teories matemàtiques: la lògica i l'àlgebra de conjunts [5]. A partir de llavors el model relacional ha evolucionat fins a aconseguir una gran solidesa, amb Date [23, 24, 26] com a principal divulgador. Quan els SGBD es basen en el model relacional s'anomenen sistemes de gestió de bases de dades relacionals (SGBDR).

El model relacional defineix el nucli dels SGBD en tres parts:

- Estructural: les relacions com a estructura principal per a representar les dades. Els tipus de dades també són necessaris per a representar les dades però no es defineixen en l'estructura principal sinó que són considerats ortogonals, cosa que a l'apartat següent expliquem en més detall.
- Manipulació: operacions sobre les relacions i que resulten en noves relacions. Són definides dualment a partir de l'àlgebra de conjunts i de la lògica, anomenades àlgebra relacional i càlcul lògic respectivament. Totes dues tenen una definició independent però són equivalents.
- Integritat: regles d'integritat o restriccions que han de complir les variables relació. La integritat es basa en aplicar operacions que han de retornar el valor cert. La integritat s'ha de complir sempre, normalment es comprova durant les assignacions. Per exemple la clau primària [23] és una regla d'integritat.

La definició de les relacions inicialment es basava en el concepte matemàtic homònim en el sentit de producte cartesià de conjunts, però el model relacional ha anat evolucionat i ara ja no són exactament el mateix. Tampoc no s'ha de confondre el terme relació (*relation*) del model relacional amb el terme relació de parentiu (*relationship*). El primer és el concepte basat en conjunts que definim a continuació mentre que el segon és el concepte de parentiu entre entitats: una a molts, molts a molts, etc. De fet, les estructures de parentiu és una de les moltes dades que

r_1

nom	edat
a	21
b	23

Figura 2.2.: Visualització com a taula d'una relació

poden ser expressades en el model relacional. Alguns autors del model relacional prefereixen el terme taula relacional (*relational table*) en comptes de relació [101].

Les relacions es defineixen com una parella de capçalera (*heading*) i cos (*body*). El cos és un conjunt de tuples on cada tuple és un conjunt de parelles atribut i valor. Tots els tuples d'una mateixa relació tenen els mateixos atributs, així es distingeix entre la capçalera de la relació –els atributs– i el cos de la relació –els tuples.

Per exemple una relació entre un nom i una edat és $r_1 = (\{\text{nom}, \text{edat}\}, \{\{(\text{nom}, \text{a}), (\text{edat}, 21)\}, \{(\text{nom}, \text{b}), (\text{edat}, 23)\}\})$. Simplificant i sense explicitar que els atributs no tenen ordre, la mateixa relació es pot expressar de forma més compacta $r_1 = ((\text{nom}, \text{edat}), \{(\text{a}, 21), (\text{b}, 23)\})$. En la definició estructural del model relacional, els valors sempre pertanyen a un tipus de dades i cada atribut és restringit a un únic tipus de dades. Així, de manera més completa hauríem d'escriure la capçalera de la relació $r1$ com $\{\text{nom} : \text{text}, \text{edat} : \text{enter}\}$.

En el context lògic del model relacional, les relacions tenen la interpretació següent: les capçaleres són predicats i els tuples són proposicions certes per al predicat. En un context informàtic també es pot interpretar que les capçaleres són funcions amb paràmetres i els tuples contenen els arguments que fan certes les funcions. Aquesta interpretació lògica és la que realment estableix el significat de les relacions en un context determinat. Així, per exemple, la capçalera de la relació r_1 podria correspondre al predicat "L'estudiant *nom* té *edat* anys" i les proposicions certes són: "L'estudiant a té 21 anys" i "L'estudiant b té 23 anys". Les relacions es defineixen segons el principi de *Closed World Assumption*; és a dir que els tuples que apareixen són proposicions certes i els que no apareixen són proposicions falses. Així, per exemple, podem dir que la proposició "L'estudiant a té 22 anys" és falsa.

Les relacions es poden representar gràficament com a taules, per exemple a la figura 2.2 es visualitza la relació r_1. Així, els conceptes de taules s'associen als de relacions i, informalment, les relacions s'anomenen taules, els tuples, files o registres, i els atributs, columnes o camps.

El nombre de tuples d'una relació s'anomena cardinal (*cardinality*) i el nombre d'atributs, grau (*degree*). Així doncs, la relació r_1 té cardinal 2 i grau 2. Hi ha només dues relacions que tenen grau zero. Tenen un nom específic, són la relació amb la capçalera i el cos buits $\text{table\_DUM} = (\{\}, \{\})$ i la relació amb la capçalera buida i un tuple buit $\text{table\_DEE} = (\{\}, \{\{\}\})$. Aquestes, però, no tenen una representació clara com a taula.

Pel que fa als operadors, en el cas de l'àlgebra relacional estan fortament relacionats amb l'àlgebra de conjunts. Així hi ha els operadors habituals de conjunts, com per exemple la unió, la diferència, la intersecció o el producte; i altres d'específics per a les relacions, com per exemple la projecció, la selecció, la junció o el reanomena [23, cap. 7]. En el cas del càlcul lògic, també anomenat càlcul relacional, estan relacionats amb la lògica de predicats. Així per exemple hi ha un operador de rang per recórrer el conjunt de tuples, el quantificador existencial o el quantificador universal [23, cap. 8].

En el model relacional es distingeix entre les variables relació o relvar (*relation variable*) i els valors relació (*relation value*). El valor relació s'anomena simplement relació, com ja hem definit fins ara. Una relvar és una variable a les qual s'assigna una relació. L'assignació a les relvar es defineix amb el símbol := a diferència de la igualtat algebraica o la definició de variables algebraiques de símbol =, com ja hem usat. Les relvar són els objectes bàsics d'emmagatzematge a les bases de dades i per tant són els objectes bàsics als quals s'apliquen les regles d'integritat. També hi ha operacions que treballen sobre les relvar, per exemple la inserció, l'actualització o l'esborrat, que són àlies de combinacions de l'assignació amb altres operacions relacionals. Unes relvar especials són les vistes. Les vistes són relvar derivades a partir d'operacions a altres relvar; és a dir són àlies d'expressions relacionals i actuen com a relvar en altres expressions. A causa d'això les vistes també s'anomenen relvar virtuals mentre que les relvar que no són derivades s'anomenen relvar base. Les relvar s'emmagatzemen en una relació especial de les bases de dades: el catàleg.

Extensió del model amb nous tipus

El model relacional ha evolucionat però no es considera que hi hagi hagut cap revolució des de la seva aparició [24, cap. 19]. Consideren que el model relacional és bastant complet i que segueix evolucionant en la comprensió de les teories i els conceptes que hi intervenen, com per exemple la recent àlgebra relacional 'A' [28, ap. A]. En aquest context d'evolució, es preveuen les investigacions que poden estendre el model relacional. Aquestes investigacions estudien propietats de les dades, com per exemple seguretat, redundància o optimitzacions de les consultes, a partir del nucli del model relacional i permeten aconseguir abstraccions més generals de les dades [24, cap. 25].

En el sentit d'extensió cal destacar la definició de nous tipus de dades, els quals estenen els SGBD en funcionalitat. Els tipus de dades (*data type*, també anomenats dominis, tipus de dades abstracte o solament tipus, són la definició d'un conjunt de valors. Cada tipus té associat un conjunt d'operadors, en alguns casos fins i tot s'entén que la definició tipus inclou aquests operadors. De manera informal [23] fa correspondre els conceptes de tipus i de relacions amb els conceptes lingüístics de noms i frases.

Com s'ha dit anteriorment, la teoria de tipus i el model relacional són ortogonals: el model relacional requereix que hi hagi un sistema de tipus de dades però diu molt poc de la naturalesa d'aquest sistema, si bé el model relacional defineix que com a mínim hi ha d'haver el tipus booleà i el tipus relació [30]. Pel que fa a implementar els tipus de dades en els SGBD, destaquen les primeres propostes fetes per Stonebraker [116] per tal que els usuaris puguin definir els seus propis tipus de dades i les de Seshadri [111] que estudia la definició de tipus de dades complexos per tal que es puguin tractar eficientment.

Normalment els SGBD tenen uns tipus predefinits, com els enters, els reals o els caràcters. Això no obstant, els tipus de dades poden definir qualssevol nous valors, com per exemple matrius, documents de text, imatges o fins i tot relacions. Aquests nous tipus de dades poden afegir estructures i operadors que ja siguin expressables amb l'àlgebra relacional o bé també poden definir-se a partir de l'àlgebra relacional. No obstant això, disposar d'un bon model d'un tipus de dades serveix per augmentar el nivell d'abstracció en el tractament dels conceptes relacionats amb aquestes dades [33].

El tipus de dades d'una relació és determinat per la seva capçalera. Així doncs, la relació $r1$ és de tipus relació $\{\text{nom} : \text{text}, \text{edat} : \text{enter}\}$. El tipus relació pot ser usat a qualsevol definició on puguin ser usats els altres tipus de dades: definicions de variables, operadors, nous tipus de dades, etc. Tot i així la definició del tipus relació és molt rígida quant als tipus dels seus atributs, cosa que, per exemple, no permet definir nous operadors genèrics per a qualsevol relació. Recentment ha aparegut una proposta preliminar de Darwen [20] per a solucionar aquest problema. Aquesta proposta permetria definir capçaleres genèriques de relacions mitjançant el símbol asterisc; és a dir capçaleres amb atributs i tipus genèrics. Així per exemple permetria usar el tipus relació $\{*\}$ que determinaria una relació de qualsevol tipus; és a dir que el conjunt de valors del tipus relació $\{*\}$ contindria totes les relacions possibles: la table_DUM, la table_DEE, la $r1$, etc. Això no obstant, encara cal flexibilitzar més les definicions dels tipus relació. Per exemple no és possible definir nous tipus de dades que siguin subtipus del tipus relació, cosa que permetria que els valors d'aquests subtipus funcionessin com a arguments en els operadors predefinits de l'àlgebra relacional.

En algunes extensions, el model relacional ha incorporat conceptes d'altres disciplines. En destaca sobretot la incorporació de conceptes dels models d'orientació a objectes en el cas dels tipus de dades i de l'herència. Des d'aquesta perspectiva, els SGBDR també s'anomenen SGBD objecte/relacionals (*object/relational*) [25]. Tot i així, Date [24, cap. 6] avisa de l'ús de la mateixa terminologia amb significat lleugerament diferent entre el model relacional i l'orientació a objectes, sobretot pel que fa als termes valor, variable i tipus. Les diferències provenen principalment del fet que el model relacional és un model de dades i el model d'orientació a objectes és un paradigma de programació i és més proper a un model d'emmagatzematge.

Casos d'exemple Finalment, com a exemple de nous tipus de dades, cal destacar que en l'àmbit dels sistemes d'informació geogràfica (SIG) hi va haver un gran impacte quan els SGBDR es van estendre amb el tipus de dades espacials [95]. Inicialment els SIG usaven programes informàtics específics per a la gestió de les dades geogràfiques. Hi va haver una recerca intensa en els SGBDR per tal d'establir models que permetessin emmagatzemar i consultar dades geomètriques i les propietats espacials de la informació geogràfica. D'aquesta manera es van definir el que es coneix com a bases de dades espacials o geoespacials, les quals han permès desenvolupar tots els avantatges dels SGBD en els SIG.

Una altra extensió de tipus important en els SGBDR s'ha produït amb l'estudi de les dades temporals [33]. Amb el model de dades temporals basat en el model relacional s'obtenen SGBDR capaços d'emmagatzemar i consultar dades històriques, que a vegades també s'anomenen multiversió perquè emmagatzemen diverses versions d'unes mateixes dades. Cal precisar, però, que de forma genèrica s'anomena dades temporals al que hem definit com a dades bitemporals a la secció 2.1, tot i que fins i tot seria més còmode anomenar-lo model d'intervals temporals per a distingir-lo clarament del model de sèries temporals. A continuació, aclarim els termes temporals, bitemporals i intervals temporals.

El problema de les dades temporals ha sigut una controvèrsia en l'àmbit dels SGBDR però finalment se n'han consolidat els conceptes [62, 63, 122]. Date, Darwen i Lorentzos [33] descriuen el model de dades temporals fortament basat en la teoria del model relacional, de fet Date [24, cap. 28] compara aquest model per dades temporals amb altres aproximacions que s'han fet no basades en el model relacional. El model de les dades temporals [33] es basa en estendre les relacions amb un atribut que és un interval temporal. Aquest interval temporal indica el rang temporal de validesa de les proposicions descrites en la relació. A més, el model també defineix les operacions necessàries per a tractar amb les dades temporals, aquestes operacions són extensions de l'àlgebra relacional. Més particularment, el model de dades temporals representa les dades històriques amb dos atributs d'interval temporal; aleshores també es poden anomenar dades bitemporals. En aquest model de dades bitemporals [33, 62, cap. 15] un dels atributs d'interval temporal s'anomena temps vàlid (*valid time*) i l'altre, temps de transacció (*transaction time*). De forma resumida, el temps vàlid registra els canvis de les proposicions en la realitat i el temps de transacció, en la base de dades.

Segons Schmidt et al. [109], els SGBD per a dades bitemporals no es consideren adequats com a SGBD per sèries temporals ja que els primers estan pensats per a històrics, es descriuen amb intervals temporals, i els segons per a anàlisi d'observacions seqüencials, es descriuen amb instants temporals. Tot i així, per a desenvolupaments futurs, observen que hi aspectes temporals comuns entre els dos sistemes i es pregunten si es podran trobar sistemes que els englobin a tots dos o cadascun necessitarà sistemes específics. A causa de la relació entre les dades temporals i les sèries temporals, sobretot pel que fa a l'atribut de temps, compararem ambdues

dades amb més detall un cop definida l'estructura de les sèries temporals (v. apartat 4.3.1).

Implementacions relacionals

Les implementacions més populars de SGBDR són les que ofereixen a l'usuari el llenguatge SQL per a operar amb les bases de dades, per exemple MySQL o PostgreSQL, a continuació ens hi referim com a SGBD SQL.

Segons Date i Darwen [29] els SGBD SQL es desvien considerablement del model relacional: permeten files duplicades, tenen ordre en les columnes, permeten valors nuls [21], etc. Les diferències entre els SGBD SQL i el model relacional han contribuït que hi hagi hagut diversos malentesos i errors, alguns dels quals han estat avaluats i desmentits [24, 101].

Actualment Date i Darwen [30] estan treballant en el *Third Manifesto* com a proposta per a obtenir SGBDR purament relacionals. Destaquen que, en el model relacional, els tipus de dades i les relacions són necessaris i suficients per representar qualssevol dades a nivell lògic. Defineixen dos principis bàsics dels SGBDR: l'*Information Principle* o *The Principle of Uniform Representation* [26], segons el qual una base de dades només conté variables relacions, i el principi d'ortogonalitat entre la teoria de tipus i el model relacional [24, cap. 6], segons el qual relacions i tipus de dades són independents i a meś els atributs de les relacions admeten qualsevol tipus.

En la proposta per a obtenir SGBDR purament relacionals Date i Darwen [28, 32] classifiquen com a *D* els llenguatges que segueixin els principis del Third Manifesto. Particularment, defineixen un llenguatge que compleix amb els principis D anomenat *Tutorial D*, amb l'objectiu que s'usi pels estudis del model relacional a nivell acadèmic. Utilitzen aquest llenguatge en les seves obres per a exemplificar els conceptes del model relacional, tot i que en alguns casos també ofereixen exemples amb SQL.

Date [24, cap. 2] considera que actualment no hi ha cap implementació comercial, per exemple les de SGBD SQL, que segueixi fidelment el model relacional. Si bé recull diverses implementacions que s'estan desenvolupant i que segueixen les especificacions del Third Manifesto [31, Projects]. Algunes d'aquestes implementacions són:

- Rel [129], que és un dels més consolidats i usa Tutorial D com a llenguatge
- Dee [52], que és una implementació molt diferent a les altres ja que defineix el llenguatge relacional com una extensió de Python

2.2.2. Recerca actual

Actualment es poden distingir quatre corrents majoritaris d'opinió en l'àmbit dels SGBD: els SGBD SQL, el NoSQL, el Third Manifesto i el NewSQL.

Els SGBD SQL són els productes tradicionals que van implementar el model relacional inicial. Aquests productes estan molt consolidats i són els més usats. Algunes propietats que tenen són les següents: garantia de propietats ACID (*atomicity, consistency, isolation, durability*) amb transaccions, optimització de consultes, emmagatzematge de grans volums de dades o gestió de seguretat i permisos. Això no obstant, ja no són considerats com l'única solució per a tots els problemes de bases dades, cosa que se sentencia amb el lema *one size does not fit all* [119, 121]; és a dir que en cada camp d'aplicació les bases de dades tenen uns requisits diferents i per tant una determinada implementació pot ser eficient en un camp concret però no en tots. A més, cada camp d'aplicació pot requerir uns requisits diferents d'eficiència: temps d'execució ràpid, poca despesa d'energia, poques dades transmeses per la xarxa, etc. Per exemple en els sistemes encastats potser no es pot implementar tot el model de dades sinó que només una part [108].

El NoSQL és un corrent modern en l'àmbit dels SGBD que té com a objectiu millorar les limitacions d'eficiència dels SGBD SQL [46, 117]. El terme NoSQL és un nom propi lexicalitzat per al lema del corrent *Not Only SQL*. Tot i que hi ha molta varietat en els productes NoSQL, s'ha d'entendre NoSQL com una crítica a les implementacions comercials actuals del model relacional, principalment els SGBD SQL, i no com una crítica al model relacional ja que els objectius parlen de millorar el rendiment dels SGBD, cosa que és només atribuïble al nivell físic però no al nivell lògic. Alguns investigadors de SGBDR veuen compatible la coexistència dels SGBD SQL amb els NoSQL perquè tenen objectius molt diferents [5]. En el corrent NoSQL hi ha molts productes diferents i s'agrupen segons el model que utilitzen, alguns exemples de models NoSQL són [46]: grafs, objectes, clau/valor, columna o semiestructurat en arbres mitjançant *Extensible Markup Language* (XML). Per exemple ZODB [48] és un sistema amb model d'objectes o Hadoop [124] té model de columna i a més és basa en un model de programació paral·lela de les consultes anomenat MapReduce [34].

El Third Manifesto, que ja hem comentat, principalment defineix els requisits *D* per a les implementacions. Defensa que els models teòrics dels SGBD tenen més sentit que mai ja que permeten mantenir una definició comuna per a les diverses implementacions que hi pugui haver. És un corrent molt crític amb els SGBD SQL i els NoSQL perquè s'allunyen del model relacional teòric [29], sobretot pel que fa al concepte de l'*Information Principle* [24, part 7]. Consideren que alguns models del NoSQL recuperen models obsolets, com el jeràrquic o el de xarxa, ja explorats en el passat; en aquest sentit avaluen alguns productes NoSQL com els SGBD XML basats en estructures d'arbre [24, cap. 14] [23, cap. 27] o els ODMG basats en objectes [24, cap. 27]. Date [24, cap. 21–25] considera que els nous models de

SGBD, a vegades anomenats post-relacionals, no estan fonamentats tan sòlidament en teories matemàtiques i la lògica de predicats com el model relacional ni tenen el mateix nivell teòric de formalització. Considera, però, la possibilitat que es pugui definir un model més potent que el relacional tot i que no veu cap indici que sigui el cas dels nous model proposats. Per tant, aconsella que per ara els SGBD no s'allunyin del model relacional. Concretament, en el Third Manifesto destaca la proposta del llenguatge Tutorial D, fidel al model relacional i que a més no és SQL. Tot i així, actualment no hi ha completada cap implementació totalment fidedigna al model relacional del Third Manifesto. Aquest és un model matemàtic de gran potència i molta abstracció i per tant és difícil obtenir-ne una implementació completa. A més les implementacions que s'estan desenvolupant són per a usos acadèmics, encara no hi ha cap intent per a aconseguir implementacions productives o comercials que tinguin en compte aspectes d'eficiència.

El NewSQL és el corrent més recent i apareix com a contrapartida dels SGBD SQL tradicionals i el NoSQL. Si bé critiquen els SGBD SQL actuals per voler ser *one size fits all*, proposen el disseny de noves implementacions dels SGBD SQL que tinguin en compte els requisits d'eficiència dels problemes actuals [117, 121]. És un corrent crític amb alguns productes NoSQL perquè no solucionen cap problema que no estigui previst en els SGBD actuals i perquè defineixen models de programació que tenen poc a veure amb la teoria de SGBD, en canvi veuen amb bons ulls la recerca NoSQL en camps on els SGBD SQL no hi funcionen gaire bé com la gestió de documents o dades semiestructurades [118]. Per exemple, remarquen que les aportacions del model de programació MapReduce ja estan suportades pels SGBD paral·lels des de fa molt temps [102]. Alguns exemples de sistemes NewSQL són: SciDB [120], els autors del qual critiquen que els classifiquen incorrectament com a NoSQL, o H-Store [10], el qual és un SGBD amb programació paral·lela.

En conclusió, per una banda el Third Manifesto considera que molts dels productes NoSQL retornen a models pre-relacional fallits. Per altra banda, els NoSQL reclamen que fins a l'arribada dels seus productes no hi havia cap sistema de bases de dades capaç de resoldre eficientment determinades aplicacions. Si bé és cert que els productes NoSQL són implementacions amb models propers al nivell físic, gràcies a això disposen de més facilitats per a investigar noves estructures eficients ja que no han de tenir en compte tota la potència del model relacional. Per una tercera banda, el NewSQL recull aquestes estructures que milloren l'eficiència i intenta expressar-les en el model relacional; així doncs els NoSQL motiven noves millores en els SGBD SQL, com per exemple el NoSQL de MapReduce ha esperonat el NewSQL de HStore. En els sistemes NoSQL, el model de grafs és el que parteix de més bona base teòrica i per tant és un candidat a formalitzar-se amb un nivell semblant que el model relacional, tot i així les aplicacions actuals estan restringides per a representar dades de tipus relació de parentiu (*relationship*). Finalment, cal destacar que, tot i aquests nous productes, els SGBD SQL tradicionals encara són el més usats perquè en moltes aplicacions segueixen sent l'opció més eficient. A més, el model relacional té molt prestigi acadèmic com a teoria dels sistemes d'informació.

2.3. Sistemes i projectes similars

Hi ha vàries implementacions de sistemes per a gestionar sèries temporals. Algunes són només l'aplicació d'un algoritme d'anàlisi per a un problema concret de sèries temporals però altres són més elaborades i es defineixen com a SGBD específics per a sèries temporals. En aquests secció resumim algunes aplicacions que considerem que implementem conceptes dels SGST.

Explorem l'estat de la recerca en sistemes i projectes similars a l'objectiu dels nostres models: gestionar les sèries temporals i aplicar-hi alguna tècnica, com la multiresolució, per tal de solucionar algunes de les propietats problemàtiques. Cal notar que hi ha una gran quantitat de sistemes propis de productes, sobretot lligats a la recol·lecció de dades de sensors, que gestionen algunes característiques de les dades adquirides. Ara bé, ofereixen capacitats molt restringides a l'àmbit on es dirigeixen els productes, és a dir que no són genèrics i són més aviat controladors del procés d'adquisició. Per exemple Keller [45] permet desar dades cada un cert període amb estructura d'anell, és a dir elimina les més antigues quan és ple, però només té un anell, a banda també permet detectar certs esdeveniments i emmagatzemar alguns estadístics de les dades. Aquests sistemes, però, són tancats i no s'especifica amb detall el seu funcionament ni la seva estructura i per tant són difícils d'avaluar.

Classifiquem els sistemes en quatre apartats segons la característica principal que els defineixi, tot i que no és una classificació absoluta ja que alguns en poden tenir més d'una:

- Sistemes genèrics
- Compressió i aproximació
- Processament en flux
- Emmagatzematge massiu

2.3.1. Sistemes genèrics

La recerca en dades bitemporals formalitza de forma adequada els SGBD per a poder tractar històrics i esdeveniments temporals [33, 62]. Això no obstant, com ja hem notat, les sèries temporals i les dades bitemporals no són exactament el mateix i no poden ser tractats de la mateixa manera [109]. Hi ha, però, certes similituds que es poden tenir en compte, per exemple les nocions de temps discret. A més, formalitzarem les sèries temporals de manera similar a com les dades bitemporals es formalitzen en els SGBDR.

Per altra banda, alguns autors descriuen sistemes genèrics per a tractar sèries temporals, és a dir amb un model adequat per a sèries temporals però sense cap tècnica específica per a processar-les. A continuació en descrivim alguns breument.

TDM Segev i Shoshani [110] presenten un model, que anomenen *Temporal Data Management* (TDM), per a dades temporals amb un llenguatge molt semblant a SQL. Les seqüències temporals que presenten són similars a les que definim en el model de SGST, inclouen la noció de regularitat i representació temporal, tot i que molt lligades a un tercer atribut que indica l'objecte de referència. Principalment estudien les operacions d'agregació sobre les sèries temporals.

Calanda Dreyer, Dittrich i Schmidt [42] proposen els requeriments de propòsit específic que han de complir els SGST i basen el model en quatre elements estructurals bàsics: esdeveniments, sèries temporals, grups i metadades, a banda de les bases de dades per sèries temporals. Implementen un SGST anomenat Calanda [41, 43, 44] que té operacions de calendari, pot agrupar sèries temporals i respondre consultes simples i ho exemplifiquen amb dades econòmiques. A [109] es compara Calanda amb els SGBD per a dades bitemporals.

Pandas Pandas [105] és una eina d'anàlisis de dades. Tot i no ser un SGBD sí que en té una forta orientació ja que gestiona les dades a partir d'una estructura tabular i amb molts conceptes relacionals. Una de les principals aplicacions és en les sèries temporals, inclou per exemple la regularització de sèries temporals. Així, Pandas és semblant a altres eines d'anàlisi estadística per a computació científica però incorpora la gestió de sèries temporals i dades similars. Un sistema similar d'anàlisi de sèries temporals, scikits.timeseries [55], s'havia desenvolupat anteriorment però actualment està previst que s'incorpori a Pandas.

2.3.2. Tècniques de compressió i aproximació

Els SGST han de gestionar les propietats problemàtiques de les sèries temporals, com les descrites a la secció 2.1.2. Principalment, el gran volum de dades comporta que s'explorin tècniques de compressió de les dades o de treballar amb dades que s'aproximin a la informació original. La compressió i aproximació es pot explorar tant amb emmagatzematge amb pèrdua de les dades originals o sense pèrdua o fins i tot intentant resoldre el problema de trobar el compromís entre les mínimes dades que poden reconstruir el senyal original amb el mínim error. A continuació descrivim breument els projectes que exploren la compressió i aproximació de sèries temporals.

T-Time Aßfalg [3] mostra un sistema que pot cercar similituds entre sèries temporals, calculades segons funcions de distàncies entre sèries temporals. Principalment, dues sèries temporals es marquen com a similars si la seva distància és menor a un llindar per cada interval de temps. A partir d'aquest mètode dissenya algoritmes eficients que implementa en un programa anomenat T-Time [4].

iSAX Camerra et al. [12] i Shieh i Keogh [114] estudien l'anàlisi i l'indexat de col·lecions massives de sèries temporals. Descriuen que el problema principal del tractament rau en l'indexat de les sèries temporals i proposen mètodes per calcular-lo de manera eficient. El mètode principal que proposen està basat en l'aproximació a trossos de la sèrie temporal [68]. Ho implementen en una estructura de gestió de dades que anomenen *indexable Symbolic Aggregate approXimation* (iSAX) [65]. Les representacions de sèries temporals que s'obtenen amb aquesta eina permeten reduir l'espai emmagatzemat i indexar tant bé com altres mètodes de representació més complexos. Aquestes tècniques de compressió són candidates per a ser usades com a funcions d'agregació d'atributs en el model de SGSTM que definim, així seria interessant poder definir agregacions en el domini freqüencial de les sèries temporals.

RRDtool [96, 97] desenvolupa un SGBD anomenat RRDtool que és molt usat per la comunitat de programari lliure en l'àmbit dels sistemes de monitoratge. A causa d'això es focalitza en unes dades en particular, les magnituds i els comptadors, i hi manquen operacions genèriques de sèries temporals. La principal característica és l'emmagatzematge de les dades amb la tècnica que anomenen Round Robin, la qual consisteix en emmagatzemar més resolució per als temps recents i en perdre resolució per als temps més antics tot gestionant els registres d'emmagatzematge de manera circular.

Hi ha diversos projectes que utilitzen RRDtool com a SGBD, en els quals hi ha sistemes de monitoratge professionals, també en l'àmbit de programari lliure, com Nagios/Icinga [58, 94] o el Multi Router Traffic Grapher (MRTG) [98]. Aquests monitors transfereixen a RRDtool la responsabilitat de gestionar l'emmagatzematge i d'operar amb les dades, i així es poden centrar en l'adquisició de dades i la gestió d'alarmes. Altres projectes adapten la tècnica de RRDtool en altres llenguatges, com per exemple JRobin [89]. També és destacable l'ús emergent de RRDtool en entorns d'experimentació, com és el cas de Zhang i Figueiredo [138] i Chilingaryan et al. [19] que hi emmagatzemen dades experimentals per posteriorment predir o validar-les. A causa del gran ús que es fa de RRDtool, sobretot en la comunitat de programari lliure, ens ha inspirat per a desenvolupar un model a partir de les principals característiques, la qual cosa és el que anomenem multiresolució.

En l'evolució de RRDtool hi ha dues millores destacables. En primer lloc, Oetiker [96] va separar el sistema de gestió de RRDtool d'un sistema de monitoratge particular, MRTG, i el va dissenyar amb l'estructura característica de Round Robin. En segon lloc, Brutlag [11] va estendre RRDtool amb algoritmes de predicció i detecció de comportaments aberrants. Actualment, s'està estudiant l'eficiència i rapidesa de RRDtool en processar les sèries temporals. RRDtool pot emmagatzemar múltiples resolucions de les dades, però Plonka, Gupta i Carder [103] troben limitacions de rendiment quan s'han d'emmagatzemar grans quantitats de sèries temporals diferents. Una solució

que observen per a aquest problema és l'aplicació de *cache* dissenyada per Carder [15], anomenada RRDcached, que permet fer funcionar simultàniament sistemes amb grans quantitats de bases de dades RRDtool.

Whisper Una eina per a visualitzar gràfics de dades que tenen forma de sèries temporals és Graphite [125]. Graphite utilitza un SGBD anomenat Whisper que té un disseny molt similar a RRDtool, de fet inicialment Graphite usava RRDtool com a sistema d'emmagatzematge.

Tsdb Deri, Mainardi i Fusco [37] desenvolupen Tsdb, un SGST d'emmagatzematge amb compressió sense pèrdua per a les sèrie temporals. Les sèries temporals han de compartir exactament els mateixos instants de temps d'adquisició i aleshores tots els valors s'emmagatzemen agrupats per temps en comptes de tenir cada sèrie temporal aïllada. Així doncs, assumeixen que les sèries temporals són regular i tenen el mateix patró de mostreig. Els valors s'emmagatzemen aplicant tècniques de compressió sense pèrdua, a diferència d'altres sistemes que també emmagatzemen tota la sèrie temporal original però amb tècniques massives, com per exemple OpenTSDB del qual comenten que té una arquitectura massa complicada i només és útil per a sistemes distribuïts.

Comparen el rendiment de Tsdb amb RRDtool i un producte SQL. Gràcies a l'estructura de Tsdb aconsegueixen un millor temps d'addició de les mesures però un pitjor temps de recuperació de les dades ja que per obtenir una sèrie temporal s'han de reagrupar els valors. Tot i així, és una aproximació interessant per a ser aplicada en els SGSTM quan cal agrupar sèries temporals que comparteixen els mateixos instants d'adquisició: aleshores es podria dissenyar una implementació amb aquesta arquitectura en què els valors fossin vectors i les operacions es processessin alhora per a totes les sèries temporals en el mateix moment de l'addició.

Emmagatzematge en memòries Flash Dou et al. [38] se centren en l'àmbit de l'emmagatzematge de sèries temporals en memòries de tipus Flash, de les quals noten que tenen propietats diferents a l'emmagatzematge tradicional en discs. Proposen emmagatzemar informació de cada sèrie temporal per a poder resoldre tres tipus de consultes: agregacions temporals, històrics basats en mostrejos aleatoris i cerca de patrons similars. La tècnica d'agregacions temporals que utilitzen és molt semblant a la de RRDtool, és a dir agregar i emmagatzemar les dades amb diferents resolucions, tot i que implementada i particularitzada per a les memòries Flash, amb registre i punters. Per a la cerca de patrons similars indexen les sèries temporals de manera similar als algoritmes de iSAX.

2.3.3. Processament en flux

Les sèries temporals també es tracten com a fluxos de dades (*data stream*) per tal de resoldre consultes d'agregació estadística de les dades mitjançant consultes aproximades. Com a fluxos de dades, s'exploren tècniques per a processar les consultes de forma incremental cada cop que arriba una dada nova. Cormode, Korn i Tirthapura [18] exploren tècniques d'agregació en flux per a sèries temporals que consideren donar més pes a les dades més recents, és a dir de manera molt similar a la multiresolució que proposem però només per a una resolució i per a unes funcions d'agregació determinades.

El processament en flux s'usa sobretot en l'àmbit de les xarxes de sensors, del qual a continuació en descrivim alguns projectes.

Cougar Fung, Sun i Gehrke [50] i Gehrke et al. [54] proposen Cougar com un SGBD per a xarxes de sensors (*sensor database systems*). El sistema té dues estructures [9]: una per a les característiques dels sensors emmagatzemades com a taules relacionals i una altra per a les sèries temporals dels sensor emmagatzemades com a seqüències de dades. Les consultes es processen de manera distribuïda. Cada sensor és un node amb capacitat de processament que pot resoldre una part de la consulta i fusionar-la amb les altres. D'aquesta manera les dades s'emmagatzemen distribuïdes en els sensors i les consultes es resolen combinant les dades amb orientació de flux, cosa que millora el rendiment del processament i es minimitza l'ús de les comunicacions. Això no obstant, l'estructura i l'estratègia de comunicació dels nodes esdevé una part crítica a configurar en aquests sistemes [36].

TinyDB Un altre prototip de SGBD per a xarxes de sensors desenvolupat paral·lelament a Cougar és TinyDB [87, 88]. A més de les característiques descrites per Cougar, aquest sistema s'implica i modifica el procés d'adquisició de les dades com ara els instants de temps, la freqüència o l'ordre de mostreig. Per exemple donada una consulta que vol correlacionar les dades de dos sensors diferents, el sistema indica als sensors implicats que han d'adquirir amb la mateixa freqüència.

2.3.4. Emmagatzematge massiu

Hi ha sistemes que aborden l'emmagatzematge massiu de les sèries temporals, és a dir de grans volums de dades, seguint l'enfocament de les VLDB. A continuació en descrivim alguns projectes.

TSDS Weigel et al. [130] noten la necessitat de mostrar les dades en tot el seu rang temporal i no només en un subconjunt com molts altres sistemes ofereixen. Desenvolupen el paquet informàtic *Time Series Data Server* (TSDS)

[131] en què es poden introduir les dades de sèries temporals per posteriorment consultar-les per rangs temporals o aplicant-hi filtres i operacions. La particularitat de TSDS és el fet que incorpora un sistema de *cache* per a les consultes que, de forma similar a la tècnica descrita per RRDtool, emmagatzema els resultats de les consultes segons la resolució i agregació realitzada. D'aquesta manera els resultats es poden aprofitar per a altres consultes similars. Això no obstant, aquestes consultes s'han de basar en els operadors predefinits de TSDS.

SciDB Stonebraker et al. [120] estudien l'emmagatzematge de dades científiques en SGBD basats en models de matrius. Les sèries temporals són les dades científiques per excel·lència, i per tant són les aplicacions que principalment exploren. Dissenyen SciDB, un SGBD que implementa les sèries temporals com a matrius i permet aconseguir anàlisis multidimensonals amb més bon rendiment. Les altres dades que acompanyen les sèries temporals les emmagatzemen en taules. Això no obstant, la diferència entre taules i matrius sembla massa del nivell físic i comporta ambigüitat per a representar les sèries temporals.

SciQL Kersten et al. [72] i Zhang et al. [139] descriuen SciQL, un llenguatge per a SGBD de dades científiques basades en matrius, del qual n'estan desenvolupant un prototip [71]. És molt semblant a la proposta de SciDB, però a diferència SciQL defineix les sèries temporals com una mescla de matrius, conjunts i seqüències. A més mostren com gestionar algunes característiques de sèries temporals com per exemple la regularitat, la interpolació o les consultes de correlació.

OpenTSDB OpenTSDB [127] és un sistema d'emmagatzematge distribuït de sèries temporals. Basa l'emmagatzematge en Apache Hadoop i HBase, els quals permeten distribuir les dades ens diferents nodes. Gràcies a aquests sistemes, pot emmagatzemar totes les dades originals ja que és una estructura en què és ràpid d'escriure-hi i localitzar les dades, cal destacar que HBase crea uns índex potents de les dades i això s'aprofita per a indexar l'atribut de temps de les sèries temporals. Per a consultar les dades defineixen el concepte d'agregadors, tot i que només per a interpolacions lineals, i les operacions d'agregació es processen en el mateix moment d'executar la consulta. Així doncs, si bé pot recuperar les dades de forma molt ràpida, restringeix les consultes a intervals temporals petits per tal que les execucions siguin ràpides. Per tant, és un sistema útil sobretot per a visualitzar i comparar intervals temporals petits de diferents sèries temporals.

Part II.

Models

3. Introducció als models

En els capítols següents es dissenya un model matemàtic per a la gestió en bases de dades de la multiresolució de sèries temporals. Es defineixen els objectes que ens permeten modelar l'estructura de les dades i els operadors que s'hi poden aplicar.

La definició del model s'estructura en dos capítols:

- Un model pels sistemes de gestió de bases de dades per a sèries temporals (SGST) que defineix mesura i sèrie temporal.
- Un model pels sistemes de gestió de bases de dades per a sèries temporals multiresolució (SGSTM) que defineix buffer, disc, subsèrie resolució, sèrie temporal multiresolució i esquema de multiresolució. Aquest model es defineix a partir del model de SGST.

En aquest capítol d'introducció, resumim els objectius de la definició dels models i aclarim alguns termes i conceptes que altrament podrien resultar confusos. Primer, relacionem el concepte de model matemàtic pels SGBD de la secció 2.2 amb el model que proposem. Segon, introduïm el context i els supòsits que utilitzarem a l'hora de treballar amb les sèries temporals en els conceptes de la secció 2.1. Tercer, introduïm les característiques principals de la multiresolució i la motivació per a definir-la.

3.1. Introducció als conceptes de model

Què és una base de dades i un sistema que la gestiona. Una definició més particular de base de dades que l'exposada en la secció 2.2 és "conjunt de dades organitzades segons una estructura coherent i accessibles des de més d'un programa o aplicació, de manera que qualsevol d'aquestes dades pot ésser extreta del conjunt i actualitzada, sense que això afecti ni l'estructura del conjunt ni les altres dades" [123, s. v. base de dades] amb la corresponent definició per als sistemes que les gestionen de "sistema informàtic que permet la gestió automàtica d'una base de dades, generalment la creació, l'emmagatzematge, la modificació i la protecció de les dades que s'hi contenen" [123, s. v. sistema de gestió de bases de dades]. En aquest cas, hem de precisar que en els capítols següents ens centrem en la teoria dels sistemes d'informació, basant-nos en el model relacional [23], per a proposar el

model lògic i deixem de banda els aspectes més d'implementació informàtica, com per exemple accedir o manipular físicament les dades.

D'aquestes definicions cal destacar la precisió en els termes d'estructura, organitzada i coherent, és a dir que s'enumeren els requisits per al correcte emmagatzematge de les dades en concordança amb l'expressat en el model relacional. En aquestes definicions, a més, caldria afegir que els SGBD han de ser capaços d'inferir informació, és a dir a partir de les dades emmagatzemades deduir-ne de noves mitjançant les consultes.

Què es modelitza amb exactitud i què amb aproximació. Els models de SGST i SGSTM que proposem són models lògics matemàtics, és a dir són models formals i abstractes que defineixen amb exactitud l'estructura i la manipulació d'unes dades independentment de la realitat. Així, definim els models com models matemàtics formals i els utilitzem com a eines simbòliques per a modelitzar la realitat.Cal no confondre la definició del model formal matemàtic amb la interpretació del model en una realitat concreta.

Tot i tenir tot el sentit matemàtic, un model lògic abstracte no té sentit pràctic si no es té en compte la descripció que fa de la realitat, és a dir que també s'ha d'interpretar el significat que té el model en la realitat. Aquest procés d'interpretació construeix un model aproximat i simplificat de la realitat; consisteix en definir, per a un determinat context, quines variables hi ha, de quin tipus són els valors, quina forma tenen, com s'interpreten, etc. Els SGBD descriuen la realitat mitjançant predicats que indiquen quins fets es consideren certs. Per tant, els models de SGBD assumeixen que aquests predicats són la realitat i deixen per a altres teories l'avaluació de com els predicats s'aproximen a la realitat. Alhora, deleguen als usuaris la interpretació del significat dels predicats a la realitat llevat d'allò que es pot expressar mitjançant regles d'integritat.

En l'àmbit dels SGST, el model assumeix com a predicat la mesura d'un valor i deixen que altres teories, com per exemple la teoria de la mesura, avaluïn fins a quin punt el valor mesurat s'aproxima a la realitat. Emfatitzant en aquest aspecte, es poden formular mesures desconegudes per a descriure errors en l'adquisició de les mesures. En el cas particular dels SGSTM, el model que es proposa descriu com a fet cert unes resolucions de la sèrie temporal i mitjançant altres teories, com per exemple la teoria de la informació, s'avalua com s'aproximen a la informació d'una sèrie temporal original.

Quin nivell es modela. En la secció 2.2 s'han nombrat els tres nivells de l'arquitectura dels SGBD: el físic, el lògic i el d'usuari. Els models que es proposen pertanyen al nivell lògic, és a dir són models lògics de l'estructura de dades i del comportament dels SGST i dels SGSTM. En el capítol 11 es proposen implementacions per a aquests models i per tant pertanyen al nivell físic. En alguns exemples i descripcions de

propietats dels models, s'avalua el significat en un context particular del model, és a dir la semàntica del model, cosa que pertany a descriure com els usuaris poden interpretar el model lògic per a modelitzar la realitat.

Àlgebra o càlcul relacional. Els models que definim són similars a l'àlgebra relacional. L'àlgebra relacional es basa en la teoria de conjunts, que és més propera a la definició d'una sèrie temporal com a conjunt de mesures i a aplicar-hi operacions de manera prescriptiva. Alternativament, el càlcul relacional, que com s'ha descrit a la secció 2.2 és equivalent a l'àlgebra relacional, es basa en la lògica de predicats i és més proper a aplicar les operacions de manera descriptiva. Això no obstant, en les definicions usem tant l'àlgebra com la lògica de conjunts segons convingui i faciliti la comprensió de les definicions.

3.2. Introducció a les sèries temporals

Una sèrie temporal és una representació per a unes variables o magnituds físiques que evolucionen al llarg del temps. En els models usarem les sèries temporals des de la visió més genèrica possible, per tant considerem una sèrie temporal com a conjunt de dades que s'han adquirit en uns certs instants de temps. En aquest sentit, les sèries temporals poden representar dades molt variades i que pertanyen a àmbits molt diferents.

La variació en el temps de magnituds com a sèries temporals són estudiades en altres teories com per exemple la teoria del senyal. L'aproximació que presentem de les sèries temporals és un raonament similar al de les altres teories però des d'un punt de vista més genèric i més propi de l'àlgebra discreta matemàtica. És a dir, les sèries temporals tenen una forma més genèrica on, per exemple, es té en compte la posició absoluta en el temps de les mostres o es pot tolerar l'inframostreig. Això no treu, però, que quan una sèrie temporal compleix amb els paràmetres de senyal digital, les operacions més adients a aplicar-hi siguin les del processament digital del senyal.

Així doncs, l'estudi genèric proposat de les sèries temporals no pretén substituir aquests estudis propis de cada àmbit sinó que pretén oferir una visió més àmplia i comuna a totes aquestes dades i complementar-lo amb aquelles dades que no tenen un comportament clarament definit. Aquest és el cas, per exemple, de les dades adquirides en monitoratges en entorns no controlats d'una variable física, on aquestes variables són aleatòries i el temps d'adquisició pot ser irregular, i per tant cal estudiar-les com a sèries temporals genèriques. Cal dir, que a vegades les sèries temporals es redueixen a seqüències temporals, és a dir a estudis de dades on només importa l'ordre en què s'han adquirit i el període d'adquisició es constant. No pretenem fer aquesta reducció sinó que tractem les sèries temporals des del punt de vista més genèric on cal saber també la posició de temps absoluta que ocupen

i la distància de temps entre els valors. Aquests estudis més particulars de les sèries temporals, els quals es focalitzen i simplifiquen algunes propietats, permeten concentrar-se més en àmbits específics i oferir solucions molt ben raonades. Per tant és interessant poder incorporar aquests estudis en els models, en aquest sentit per exemple utilitzarem conceptes de la teoria del senyal per a interpretar propietats de les sèries temporals.

Interpretació de la sèrie temporal. La interpretació genèrica d'una sèrie temporal és un conjunt de predicats 'en el temps t la variable observada té el valor v' pertanyents a una mateixa variable o fenomen físic. De forma més particular, i en una interpretació més lligada a l'adquisició i monitoratge continu de fenòmens, una sèrie temporal indica la mesura d'un valor en un temps, és a dir que el fet que es constata com a cert és que segons un rellotge i un aparell de mesura s'ha adquirit una parella de temps i valor.

El models de SGBD, com el model relacional, defineixen la metodologia per tal d'assegurar la correctesa en la inferència d'informació a partir dels fets que es donen com a certs. Alhora es donen com a falsos els fets que no són constatats; és a dir que si en una sèrie temporal hi apareix una mesura en un temps t de valor v significa alhora que és fals que en aquell instant s'ha mesurat un altre valor diferent de v. Particularment, en el cas que en una sèrie temporal no hi apareix un temps t significa que és fals que en aquell instant s'ha mesurat qualsevol valor; així, si s'ha mesurat però s'ha obtingut un valor erroni aleshores hauria d'aparèixer marcat amb un valor especial.

Atesa aquesta interpretació de valor adquirit en un instant de temps per a una mateixa variable observada, en una sèrie temporal no hi pot haver instants de temps repetits. Altrament, no tindria sentit que un mateix aparell hagués mesurat alhora dos valors diferents en el mateix instant.

Semàntica de les sèries temporals. Les sèries temporals s'utilitzen per a modelitzar aspectes de la realitat i per tant per a cada cas cal estudiar-ne l'adequació. Això no obstant, en el nostre cas ens centrem en els aspectes formals dels models, tot i que cal exposar alguns aspectes semàntics per a poder comprendre'n la utilitat. Així, cal tenir en compte dues consideracions.

Per una banda, en el context del monitoratge, per a estudiar la deducció d'informació sobre la variable mesurada a partir del procés d'adquisició s'ha de complementar amb les teories adients. Per exemple si l'aparell de mesura està avariat la informació inferida en el SGST serà certa des del punt de vista que aquell és el valor mesurat per l'aparell però, evidentment, no serà cert que la variable hagi tingut aquell valor. De fet, aquest és el principi d'inferència d'informació que segueix el model relacional de bases de dades: a partir dels fet que es donen com a certs estableixen com a certa la informació que s'infereix, els models asseguren que aquest raonament sigui

correcte, però responsabilitzen a l'usuari d'interpretar si aquella operació efectivament es correspon amb la informació que vol inferir. És a dir, els operadors només tenen en compte l'estructura de les dades mentre que el significat contextualitzat dels operadors és extern al model; com de fet és el cas d'altres àlgebres que tampoc no defineixen com s'han d'interpretar la validesa dels resultats.

Per altra banda, no totes les sèries temporals són adquirides directament, car poden ser resultat d'operacions amb altres sèries temporals o resultat de consultes on el temps sigui una variable, com per exemple consultar les mitjanes mensuals de temperatures. En aquests casos també és important no perdre de vista la interpretació de la validesa dels resultats.

Així doncs, en els models no incloem l'etapa d'adquisició ni de mostreig de les sèries temporals sinó que partim del fet que les mesures ja han estat capturades i s'ha raonat sobre l'adequació d'aquest procés. Això no obstant, cal notar que alguns SGST proposen el fet d'influir sobre el procés d'adquisició a partir de la informació gestionada [88], per exemple per poder sol·licitar d'obtenir més mostres si s'observen variacions estranyes o per reduir la freqüència de mostreig si es considera que el sistema té un estat estable.

En resum, les sèries temporals tenen un atribut, l'instant de temps, que ofereix unes particularitats a l'estructura de dades i amb què cal operar coherentment. Tant els instants de temps com els valors poden ser de qualsevol tipus, tot i així els exemplificarem amb nombres reals per tal de facilitar-ne la comprensió i per ser més propers a les anàlisi de sèries temporals basades en variables numèriques [78].

3.3. Introducció a la multiresolució

La multiresolució és una tècnica que s'aplica a una sèrie temporal per tal de compactar-ne i resumir-ne certa informació. Bàsicament la multiresolució consisteix a calcular un conjunt de resolucions d'una sèrie temporal on cada resolució consisteix en aplicar una funció d'agregació a les mesures cada cert període de temps. A més, cada resolució inclou un paràmetre per tal d'afitar el nombre de valors emmagatzemats.

La idea bàsica és utilitzar la multiresolució per a descriure sèries temporals de forma que hi hagi més resolució per a les dades més recents i menys resolució per a les dades més antigues. Tot i així, es podrien establir variacions d'aquesta estructura mitjançant les quals, per exemple, es pogués retenir una resolució per a un període temporal que ha resultat interessant o que calgués investigar més profundament. Aquesta idea de multiresolució prové de l'SGBD RRDtool [97], del qual en estudis anteriors n'hem analitzat profundament els conceptes i n'hem abstret i formalitzat les característiques essencials [81, 82]. Els objectius principals són la formalització

d'un model abstracte de SGBD per a la multiresolució i la inclusió de conceptes més genèrics per tal de descriure els SGSTM contextualitzats en els SGST.

La multiresolució aplicada a una sèrie temporal implica una selecció d'informació i per tant és alhora una compressió de dades amb pèrdua. Abans d'aplicar la multiresolució cal decidir quins atributs se seleccionaran mitjançant uns paràmetres, els quals bàsicament són les definicions del períodes de temps, les funcions que han d'agregar els atributs i el nombre màxim de valors que s'han d'emmagatzemar. Així doncs, aplicar la multiresolució ha de ser una decisió consensuada, s'ha de tenir en compte que és una compressió amb pèrdua i s'ha de pensar adequadament la configuració dels paràmetres. O, dit d'una altra manera, l'usuari ha de ser conscient que vol gestionar les sèries temporals amb multiresolució; en certa manera un sistema no pot decidir autònomament d'utilitzar la multiresolució com a equivalent a la sèrie temporal original sense avisar l'usuari. Com en els SGST, en el cas dels SGSTM l'usuari també ha d'interpretar la validesa dels resultats. En aquest cas, però, la multiresolució produeix l'efecte de compressió amb pèrdua que estudiarem amb més detall en el capítol 9.

El model que presentem d'SGSTM defineix els termes i conceptes de la multiresolució i els operadors genèrics que hi treballen. En els capítols posteriors al model d'SGSTM, comentarem aplicacions i variacions interessants dels SGSTM.

3.3.1. Característiques de la multiresolució

Un SGSTM és un SGST amb capacitats de multiresolució. A continuació resumim les característiques que els SGSTM milloren respecte als SGST.

- Gran volum de dades. Els sistemes de monitoratge adquireixen una gran quantitat de dades dels sensors. Aquestes dades contenen informació que ha de ser observada, tant en línia amb l'adquisició com en diferit, i per tal de poder processar-la cal reduir el volum de dades. Una de les característiques de la multiresolució és la selecció i l'emmagatzematge dels segments més interessants de les dades. Aquests segments són el conjunt de resolucions per a cada sèrie temporal que l'usuari pot configurar com extreure i resumir mitjançant diversos períodes de temps i funcions. En la multiresolució la mida de les sèries temporals queda afitada, cosa que és útil per a sistemes d'emmagatzematge que necessiten controlar-ne l'espai.
- Visualització. La multiresolució també és útil quan es visualitzen gràficament les sèries temporals ja que permet a l'usuari seleccionar el millor rang temporal i la resolució que s'adeqüen a la pantalla. No cal processar amb més quantitat de dades que la que realment es pot mostrar.
- Validació de dades. Els sistemes de monitoratge adquireixen dades però poden ocórrer alguns problemes que tenen efecte en el procés posterior d'anàlisi

de les sèries temporals. Un dels principals problemes ocorre quan els monitors no poden adquirir una dada, cosa que es coneix com a forats, o bé quan adquireixen una dada erròniament, com per exemple les dades atípiques o aberrants [107]. Les funcions d'agregació d'atributs de la multiresolució poden cooperar en la validació, el filtratge i la reconstrucció d'aquestes dades desconegudes per tal de conservar històrics consistents.

- Regularització de les sèries temporals. Un altre efecte secundari del monitoratge ocorre quan el període de mostreig no és constant, és a dir quan les dades resultants no estant equiespaiades en el temps. Aquestes no regularitats poden provenir de fluctuacions en el rellotge de mostrejos periòdics o bé de mostrejos no periòdics basats en esdeveniments [74]. Un dels objectius de la multiresolució és regularitzar els intervals de temps de la sèrie temporal, és a dir que les resolucions que en resulten són regulars en el temps. Aquest procés de regularització és útil per a aplicar posteriorment algoritmes d'anàlisi de sèries temporals que assumeixen que les sèries temporals són regulars, però també per a calcular altres resolucions de la sèrie temporal com per exemple observar dades periòdiques amb intervals de mes o d'any.

- Resum de la informació. La multiresolució extreu i selecciona les característiques desitjades de les dades mitjançant diverses funcions d'agregació d'atributs. Per tant, emmagatzema resumidament la informació que l'usuari posteriorment pot consultar. Això no obstant, aquesta selecció de la informació s'ha de determinar a priori considerant el context en què les consultes futures s'hauran de realitzar.

- Computació en flux. Els resums d'informació que resulten de la multiresolució són cars de computar, sobretot si s'han emmagatzemat grans quantitats de sèries temporals o bé si es té en compte que en monitoratges periòdics constantment arriben dades noves. El model de SGSTM permet calcular la multiresolució en flux amb l'adquisició de dades, és a dir al mateix temps que arriben noves mesures de la sèrie temporal computa incrementalment el nou resultat de multiresolució. D'aquesta manera el temps de còmput es pot repartir des del moment d'adquisició fins al moment de la consulta. Això, però, requereix que les mesures s'insereixin amb ordre als SGSTM.

- Computació distribuïda i paral·lela. En el cas que es vulgui computar la multiresolució en temps diferit del procés d'adquisició de dades, és a dir que es vulgui emmagatzemar tota la sèrie temporal i després computar amb totes les dades, cal una computació intensiva. A tal efecte, la multiresolució també es pot computar distribuïdament i paral·llament en diversos nodes de computació.

Tot i així, també pot ser útil de complementar els SGSTM amb les capacitats d'altres SGBD. Per una banda, es poden usar per a emmagatzemar els valors originals en qualitat de dipòsit a llarg termini en cas que calgui realitzar alguna consulta

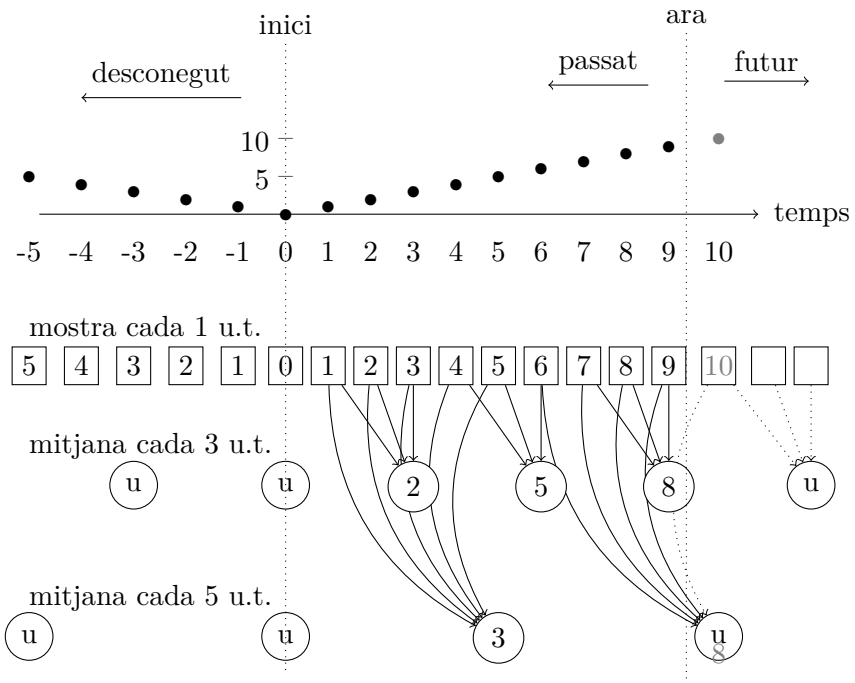


Figura 3.1.: Diagrama d'una instantània en la multiresolució d'una sèrie temporal amb mostreig regular

imprevista en temps diferit. Par altra banda, s'hi pot emmagatzemar informació relacionada amb les sèries temporals com per exemple les unitats dels valors, la localització del sensor, etiquetes de classificació, darrer valor mesurat, etc.

3.3.2. Motivació per a la multiresolució

A continuació mostrem la motivació per a la multiresolució mitjançant dos exemples: la figura 3.1 i la figura 3.2. Les figures mostren el càlcul de la multiresolució per a una sèrie temporal regular i una d'irregular, respectivament. Assumim que les figures mostren una instantània entre els instants de temps nou i deu.

A la part superior de les figures hi ha el gràfic d'una sèrie temporal en què l'eix de temps té qualssevol unitats de temps (u.t.) i l'eix de valors qualssevol unitats. És una sèrie temporal senzilla que val 1 en l'instant -1, 0 en el 0, 1 en l'1, 2 en el 2, etc. L'eix vertical *ara* indica l'instant en què s'ha pres la instantània, així que el temps anterior és el passat i el temps posterior és el futur, els esdeveniments del qual estan pintats en gris. L'eix *inici* indica l'instant zero u.t. en què el sistema ha començat a adquirir dades, així les dades anteriors són desconegudes.

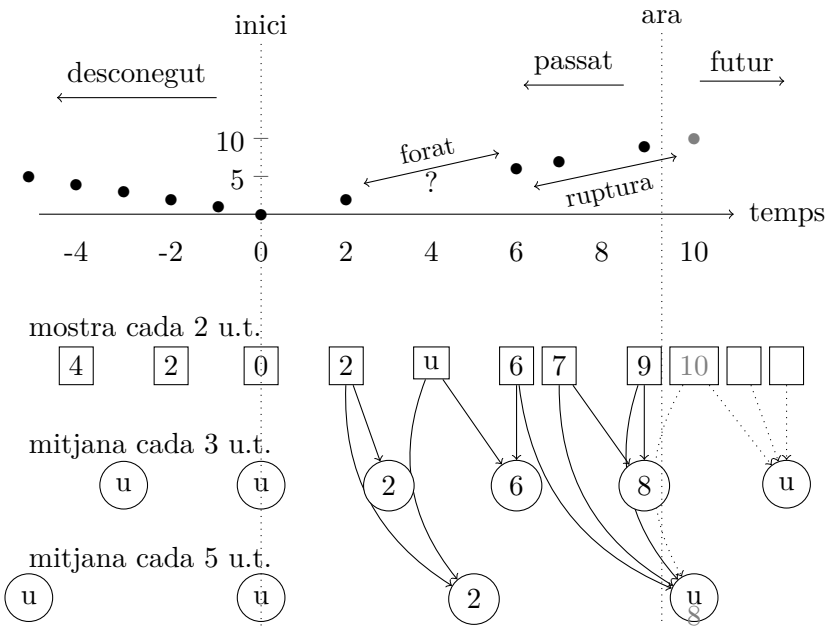


Figura 3.2.: Diagrama d'una instantània en la multiresolució d'una sèrie temporal amb mostreig irregular

A la part inferior de les figures hi ha un diagrama que mostra l'acció de la multiresolució. La primera línia mostra enquadrades els valors numèrics adquirits de la sèrie temporal. La segona i la tercera línia mostren encerclades les dades que s'emmagatzemaran a la base de dades segons un esquema de multiresolució particular que consisteix en calcular dues resolucions de la sèrie temporal: una calcula la mitjana dels valors cada tres u.t. i l'altra calcula la mitjana cada cinc u.t. En aquest cas, la mitjana és la funció que actua com a selector d'informació de la sèrie temporal mitjançant estadístics d'agregació. Totes les dades emmagatzemades abans de l'instant zero són desconegudes (u) i totes les dades futures també són marcades com a desconegudes fins que el temps avanci, tot i que en alguns casos en gris mostrem els valors que prendran.

A la figura 3.2 la sèrie temporal s'ha adquirit cada una u.t. de forma regular, és a dir que hi ha valors per a cada adquisició. Les fletxes mostren com es resumeixen les dades adquirides per tal d'emmagatzemar-les: així de cada tres mostres se n'emmagatzema la mitjana i, independentment, cada cinc mostres se n'emmagatzema una altra mitjana. Per als valors futurs, quan el temps avanci una u.t. aleshores s'adquirirà el valor 10 i es podrà calcular la mitjana de cada 5 u.t. per a l'instant 10, la qual resultarà en el valor 8, però no es podrà calcular la mitjana de cada 3 u.t. fins a l'instant 12.

A la figura 3.1 la sèrie temporal s'ha adquirit cada dues u.t. de forma irregular, és

a dir que manquen valors en alguna adquisició, cosa que es marca com un forat i un valors adquirit desconegut, o bé l'adquisició no s'ha fet exactament cada dues u.t., cosa que es marca com una ruptura. Aquests són dos exemples de possibles problemes en el monitoratge, és a dir que es volia mostrejar cada dues u.t. però per alguna raó no s'ha pogut fer en l'instant 4 i al voltant de l'instant 8 hi ha hagut una ruptura en el rellotge que ha adquirit en els instants 7 i 9. L'esquema de multiresolució emmagatzemat té la mateixa forma que en el cas regular, és a dir sense que hi afectin les irregularitats. Aquí, les fletxes mostren com es calcula la mitjana a partir dels valors adquirits, és a dir que ara no hi ha ni tres ni cinc valors disponibles per a fer la mitjana de cada resolució. Alguns valors emmagatzemats coincideixen però altres difereixen, especialment quan la mitjana opera amb valors desconeguts dels quals en aquest exemple operem sense tenir-los en compte; en els models detallarem més bé aquests casos.

4. Model SGST

En aquest capítol es defineix un model per als sistemes de gestió de bases de dades per a sèries temporals (SGST). Aquest model s'estructura en base a dos objectes principals, mesures i sèries temporals. Ambdós tenen un atribut de temps, el qual requereix un tractament adequat. El model de SGST es dissenya en tres parts.

- Primer, es defineix el model d'estructura de les dades, és a dir, la forma com es descriuen les mesures i les sèries temporals.
- Segon, es defineix el model d'operacions sobre les dades, és a dir, els operadors bàsics que permeten modelar el comportament i la manipulació de les sèries temporals.
- Tercer, es descriuen propietats de les sèries temporals. Les sèries temporals adquireixen propietats variades depenent del context on s'apliquin.

4.1. Model estructural de dades

L'estructura d'un SGST està formada per quatre conceptes principals: temps, valor, mesura i sèrie temporal. Al final d'aquesta secció, mostrem alguns exemples de sèries temporals amb valors concrets.

Una sèrie temporal és una relació de temps i valors. A cada parella temps-valor l'anomenem mesura. Així doncs, una sèrie temporal és un conjunt de mesures i una mesura es correspon amb un valor mesurat en un instant de temps.

4.1.1. Temps

El temps és la variable que permet ordenar les mesures. Anomenem *domini del temps* al conjunt $\mathcal{T}$ de tots els possibles valors de temps. $\mathcal{T}$ pot ser tant un conjunt finit com infinit i normalment serà un conjunt tancat per a poder incloure les mesures indefinides (v. definició 4.6) com a límits. Per tal de facilitar la comprensió, en aquest document assumim que $\mathcal{T}$ és el conjunt estès de nombres reals $\bar{\mathbb{R}} =$

$\mathbb{R} \cup \{+\infty, -\infty\}$ [13, 133], també anomenat recta real acabada, el qual és un conjunt tancat.

El conjunt estès de nombres reals té dos punts límits corresponents al valor impropi infinit, aleshores en notació d'interval el conjunt $\mathcal{T}$ es pot escriure com $\bar{\mathbb{R}} = [-\infty, +\infty]$. En referència amb el conjunt dels nombres reals $\mathbb{R}$, les relacions d'ordre i algunes operacions aritmètiques s'estenen al conjunt $\bar{\mathbb{R}}$ [13]. Algunes expressions esdevenen indefinides (p.ex. $0/0$) i altres depenen del context, com és el cas de l'expressió indeterminada $0 \times \infty$ que per exemple en la teoria de la mesura habitualment es defineix com $0 \times \infty = 0$ [133].

El conjunt dels reals és un espai mètric ja que té definida una funció distància (o mètrica), com per exemple la distància euclidiana. Com a conseqüència, ens permet distingir entre instants de temps (els elements del conjunt) i durades (la mètrica). Observant els instants de temps com a punts en la recta real, les durades com a segments de la recta real i especificant un instant de temps com a marc de referència, es pot definir el temps com a sistema de coordenades [39, 74]. A continuació definim el temps de manera que puguem ordenar esdeveniments, mesurar durades d'esdeveniments i establir quan esdevenen; és una aproximació ingènua sense abastar detalls complicats del concepte temps [40].

Definició 4.1 (Temps)**.** *Sigui $\mathcal{T} = \bar{\mathbb{R}}$ el domini del temps. Anomenem un element $t \in \bar{\mathbb{R}}$ com a* instant de temps*. L'element $0 \in \bar{\mathbb{R}}$ és el* marc de referència*.*

Siguin $s, t \in \bar{\mathbb{R}}$ dos instants de temps. Definim la durada de temps *entre s i t com el valor $d \in \bar{\mathbb{R}}$ que mesura la distància en unitats de temps entre tots dos instants de temps, és a dir $d = |s - t|$.*

En resum, els instants de temps es poden veure com una seqüència de valors reals que ordenen els esdeveniments i entre dos instants de temps es pot determinar una durada. El marc de referència és l'instant de temps que correspon a l'*origen* del sistema de coordenades. Expressem tant els instants de temps com les durades amb un real que té unitats de temps. Aquestes unitats són 'segons' en sistema internacional.

Estàndards de temps

Els estàndards de temps especifiquen com s'ha de mesurar el pas del temps i com s'han d'assenyalar els instants de temps. Allen [2] recull diferents estàndards de temps que existeixen, dels quals a continuació comentem els més habituals.

Actualment l'estàndard de temps habitual per mesurar el pas del temps és el temps atòmic internacional (TAI, del fr. *temps atomique international*), del qual se'n deriva un altre estàndard més conegut que és el temps universal coordinat (UTC, del fr. *temps universel coordonné* i de l'angl. *Coordinated Universal Time*). Ambdós

estàndards assenyalen els instants de temps segons el calendari gregorià i segons el dia julià. Actualment, de forma genèrica s'utilitza UTC per a sincronitzar rellotges, tot i que en el futur es podria canviar per altres estàndards nous com per exemple un anomenat Temps Internacional o simplement TI, el qual també es basaria en el TAI.

El dia julià utilitza un estàndard de comptar el temps com a nombre de dies que han passat des d'una data concreta, la qual s'anomena època. L'època es correspon amb el concepte d'instant de temps marc de referència de la definició 4.1. Per defecte l'època se situa a l'inici del Període Julià tot i que també se solen utilitzar altres dates assenyalades.

Així, un estàndard semblant al julià és l'Hora POSIX o Hora Unix, el qual compta el nombre de segons des de l'1 de gener de 1970 basant-se en les mesures d'UTC. L'Hora Unix és l'estàndard de temps habitual en els sistemes operatius de la família Unix. No obstant això, aquest estàndard presenta un problema d'ambigüitat a causa que no té en compte els segons addicionals d'UTC.

Calendari

Un cas particular del temps és el calendari. Els calendaris són definicions pel domini temps que consisteixen en noms per als punts de la línia de temps i regles per establir la durada entre ells per tal que el temps tingui certa relació amb la rotació de la Terra. A l'apartat anterior hem definit el domini temps de manera genèrica amb el conjunt de reals, els quals exemplifiquen el concepte de sistema de coordenades de temps absolut sense entrar en detall en conceptes de calendari.

Dreyer, Dittrich i Schmidt [42] situen els calendaris i les seves operacions com a essencials en els SGST. Tanmateix, pot no ser necessari modelar les dates i regles de calendari en el model de temps. Els calendaris es poden observar com a noms que fan referència a instants de temps quantificables, com els de la definició 4.1. Aleshores, només cal una eina que sigui capaç de convertir els noms de calendari a instants de temps.

El fet que un calendari sigui més o menys complicat no afecta al model de SGST, sols té incidència en les funcions de conversió d'instant de temps a calendari i viceversa. Tampoc afecta que els calendaris siguin ambigus (p.ex. dos noms per al mateix instant o instants sense nom) o que continguin propietats impredictibles (p.ex. cas dels segons addicionals en UTC) ja que aquests casos es corresponen amb la bona definició dels sistemes de calendari.

Així doncs, els calendaris en el model de SGST es poden implementar com una extensió del model de temps. El tipus de dades ordinal de calendari gregorià implementat per Date, Darwen i Lorentzos [33, cap. 16] pot servir com a guia per a la implementació dels calendaris en els SGST.

4.1.2. Valor

El valor és la variable que indica la magnitud de la dada mesurada. Per tal de no restringir el model a cap àmbit de mesura, definim genèricament el concepte de valor. Així, el valor és qualsevol element que és d'un tipus de dades, també anomenat domini i que simbolitzem amb $\mathcal{V}$; és a dir, un valor és un objecte que pertany a un determinat conjunt de valors $\mathcal{V}$ i que té associades les operacions que s'hi poden aplicar. Exemples de tipus de dades són els enters, els reals, les cadenes de text i les estructures de dades com vectors, llistes o relacions.

El valor quan forma part d'una sèrie temporal pot preveure una dada que defineixi el valor indefinit. Aquest valor indefinit és un valor impropi, és a dir no vàlid, del conjunt de valors possibles. Així cada tipus de dades pot anar associat amb un valor indefinit corresponent. Seguint l'exemple amb nombres reals per a la variable temps, en aquest document assumim que el domini $\mathcal{V}$ és el conjunt dels reals estès projectivament $\mathbb{R}^* = \mathbb{R} \cup \{\infty\}$ [14]. D'aquesta manera, com a valor indefinit en aquest document usem el valor infinit (∞), el qual és un valor impropi del conjunt dels reals. En altres sistemes, es podrien utilitzar més símbols per a precisar diferents casos de valors impropis, com per exemple assenyalar casos de valors no numèrics (NaN, *not a number*).

El valor indefinit s'utilitza per identificar que el valor d'una mesura és desconegut. Un valor és desconegut quan en el moment de fer la mesura es desconeix o és erroni. També pot esdevenir desconegut posteriorment si és marcat com a erroni o descartat després d'un processament de les dades (v. apartat 4.3.3).

4.1.3. Mesura

Una mesura és un valor mesurat en un determinat instant de temps. Per tant, és una parella de temps i valor.

Definició 4.2 (Mesura)**.** *Sigui $v \in \mathcal{V}$ un valor i $t \in \mathcal{T}$ un instant de temps. Definim una* mesura *m com el tuple $m = (t, v)$, en què v és el valor de la mesura i t és l'instant de temps en que s'ha pres aquesta mesura.*

El domini d'una mesura m, notat com a $\operatorname{dom} m$*, és el domini del seu valor.*

Sigui $m = (t, v)$ una mesura, escrivim $V(m)$ per a referir-nos a v i $T(m)$ per a referir-nos a t.

L'instant de temps indueix la relació d'ordre entre les mesures. Definim dues relacions d'ordre diferents.

Definició 4.3 (Ordre semitemporal)**.** *Siguin m i n dues mesures. Anomenem* ordre semitemporal *a la relació binària $m \leq n$ que definim com $m \leq n \iff T(m) < T(n) \vee (T(m) = T(n) \wedge V(m) = V(n))$.*

Definició 4.4 (Ordre temporal)**.** *Siguin m i n dues mesures. Anomenem* ordre temporal *a la relació binària $m \leq^t n$ que definim com $m \leq^t n \iff T(m) \leq T(n)$.*

Noteu que l'ordre semitemporal és un ordre parcial mentre que l'ordre temporal és un ordre total. Amb aquestes relacions d'ordre queden definides les operacions de comparació i igualtat entre mesures: $m < n$ i $m <^t n$, $m = n$ i $m =^t n$, $m > n$ i $m >^t n$, etc. Precisament, les dues relacions d'ordre només es diferencien en les operacions d'igualtat.

Mesures multivaluades

Les mesures poden contenir alhora més d'un fenomen mesurat quan aquests comparteixen els instants de temps de mesura; és a dir que hi ha una col·lecció de valors mesurats en el mateix instant de temps. Aleshores les mesures poden esdevenir tuples n-dimensionals $(t, v_1, \ldots, v_n)$, de manera semblant a com ho defineix Aßfalg [3]. Anomenem aquestes mesures com a *mesures multivaluades.*

Mesures indefinides

En les definicions de temps i valor s'han estès els conjunts amb valors impropis, concretament s'ha exemplificat amb el conjunt estès $\bar{\mathbb{R}} = \mathbb{R} \cup \{+\infty, -\infty\}$ pel temps i amb el $\mathbb{R}^* = \mathbb{R} \cup \{\infty\}$ pel valor. Aquesta extensió amb l'element impropi infinit (∞) dóna com a resultat unes mesures impròpies que anomenarem mesura de valor indefinit i mesura indefinida.

Definició 4.5 (Mesura de valor indefinit)**.** *Definim* mesura de valor indefinit *com el tuple (t, v) en què el valor és indefinit $v = \infty$ i l'instant de temps és qualsevol $t \in \bar{\mathbb{R}}$.*

Definició 4.6 (Mesura indefinida)**.** *Definim* mesura indefinida *com el tuple (t, v) en què el valor és qualsevol $v \in \mathbb{R}^*$ i l'instant de temps és indefinit $t \in \{+\infty, -\infty\}$.*

Així doncs, sigui m una mesura de valor indefinit, aquesta pren la forma $m = (t, \infty)$. Sigui m una mesura indefinida, aquesta pren la forma $m = (+\infty, v)$ per la positiva i $m = (-\infty, v)$ per la negativa, les quals normalment anotarem també amb valor indefinit: $m = (+\infty, \infty)$ i $m = (-\infty, \infty)$ respectivament.

4.1.4. Sèrie temporal

Una sèrie temporal és un conjunt de mesures del mateix fenomen. Com a conseqüència, en una sèrie temporal les mesures són homogènies, és a dir els temps pertanyen al mateix domini i els valors pertanyen al mateix domini. Tradicionalment s'anomenen sèries temporals tot i que alguns autors les anomenen seqüències temporals, per exemple Hetland [57].

Definició 4.7 (Sèrie temporal)**.** *Sigui $S = \{m_0, \ldots, m_k\} \subset \mathcal{T} \times \mathcal{V}$ un conjunt finit de mesures del mateix tipus. Aleshores, S és una* sèrie temporal *si i només si no hi ha temps repetits $\forall i, j : j \in [0, k] \wedge i \neq j : T(m_i) \neq T(m_j)$.*

Definim el domini d'una sèrie temporal S, notat com a $\operatorname{dom} S$*, com el domini de les seves mesures.*

Com que en una sèrie temporal no hi ha temps repetits, no hi ha discrepància en l'operació d'igualtat i per tant les mesures contingudes tenen un ordre total. Això no obstant, donades dues sèries temporals diferents, entre totes les mesures hi pot haver temps repetits i per tant l'ordre temporal i el semitemporal indueixen a les dues operacions d'igualtat com ja s'ha comentat.

Per ser un conjunt, les sèries temporals tenen mesura de cardinalitat.

Definició 4.8 (Cardinal)**.** *Sigui S una sèrie temporal, el cardinal de la sèrie temporal, notat com a $|S|$, és el nombre de mesures que conté la sèrie temporal.*

Una sèrie temporal sense mesures és la sèrie temporal buida que notem com a $\emptyset = \{\}$. És a dir que no té cap element i per tant $|\emptyset| = 0$.

Formes d'una sèrie temporal

Una sèrie temporal s'expressa com un conjunt i com a tal és susceptible d'aplicar-hi els conceptes del model relacional dels SGBDR (v. § 2.2), a continuació expressem la forma de sèrie temporal seguint també el concepte de relació. Diferenciem entre tres formes possibles d'una sèrie temporal: canònica, multivaluada i doble.

La forma bàsica d'una sèrie temporal és la de parelles de temps i valor, l'anomenem forma canònica.

Definició 4.9 (Forma canònica)**.** *Sigui $S = \{m_0, m_1, \ldots, m_k\} \subset \mathcal{T} \times \mathcal{V}$ una sèrie temporal, la forma canònica com a relació s'escriu com $S = (\{t : \mathcal{T}, v : \mathcal{V}\}, \{\{(t, t_0), (v, v_0)\}, \{(t, t_1), (v, v_1)\}, \ldots, \{(t, t_k), (v, v_k)\}\})$; és a dir és una parella amb la capçalera i el conjunt de valors certs.*

Així doncs, sigui $\emptyset = \{\}$ una sèrie temporal buida, modelada com a relació s'escriu com $\emptyset = (\{t : \mathcal{T}, v : \mathcal{V}\}, \{\})$.

S

t	v
t_0	v_0
t_1	v_1
...	...
t_k	v_k

$\emptyset$

t	v

Figura 4.1.: Visualització com a taula d'una sèrie temporal

A causa del format esquemàtic de les sèries temporals, en simplifiquem l'escriptura de la forma canònica com a conjunt de tuples (t, v) en què t és el temps i v és el valor. Així doncs quan no hi ha dubte sobre els dominis ni els noms d'atributs, una sèrie temporal es pot escriure de manera simplificada com a $S = \{(t_0, v_0), (t_1, v_1), \ldots, (t_k, v_k)\}$, la qual es correspon amb la forma de la sèrie temporal expressada inicialment a la definició 4.7.

Tal com s'utilitza en les relacions, les sèries temporals es poden visualitzar com a taules. La sèrie temporal S i la $\emptyset$ es visualitzen com a taula a la figura 4.1. Per ser un conjunt de mesures, s'observa una sèrie temporal en la forma canònica com una relació de grau dos on la capçalera conté els atributs temps i valor. Ambdós atributs tenen els dominis de temps i valor descrits a les seccions 4.1 i 4.1.2, com per exemple el tipus de dades reals estesos. Les relacions de sèries temporals inclouen algunes particularitats:

- El predicat és similar a: «En el temps t s'ha mesurat el valor v»
- Els temps no poden ser repetits: és una restricció que indica que l'atribut t és la clau primària
- Els valors mesurats han d'estar associats al mateix fenomen o fenòmens.

Les sèries temporals poden mesurar alhora més d'un fenomen quan aquests comparteixen els instants de temps de mesura; és a dir contenir mesures multivaluades. Anomenem aquestes sèries temporals com a sèries temporal multivaluades.

Definició 4.10 (Sèrie temporal multivaluada)**.** *Anomenem sèrie temporal multivaluada a una sèrie temporal que té més d'un atribut de valors. Així una sèrie temporal multivaluada té la forma simplificada* $\{m_0, m_1, \ldots, m_k\} \subset \mathcal{T} \times \mathcal{V}_1 \times \cdots \times \mathcal{V}_n$ *on cada mesura* m_i *és un tuple* $m_i = (t, v_1, \ldots, v_n)$ *on* t *és un instant de temps i* v_1, ..., v_n *són valors.*

La forma completa com a relació d'una sèrie temporal multivaluada buida és $\emptyset = (\{t : \mathcal{T}, v_1 : \mathcal{V}_1, \ldots, v_n : \mathcal{V}_n\}, \{\})$

Com ocorre en les relacions, el nom dels atributs d'una sèrie temporal pot ser decidit per l'usuari. Per exemple, una sèrie temporal multivaluada amb tres atributs anomenats és $\emptyset = (\{t : \bar{\mathbb{R}}, \text{temperatura} : \mathbb{R}^*, \text{consum} : \mathbb{R}^*, \text{volum} : \mathbb{R}^*\}, \{\})$.

Una sèrie temporal multivaluada es pot escriure en forma canònica de mesures (t, v). D'aquesta manera, les mesures m_i d'una sèrie temporal multivaluada en forma canònica són tuples $m_i = (t, (v_1, v_2, \ldots, v_n))$. Així, per al cas de la sèrie temporal multivaluada buida $\emptyset$, en forma multivaluada canònica és $\emptyset = (\{t : \mathcal{T}, v : V\}, \{\})$ on el domini de l'atribut valor és de tipus relació $V = \{v_1 : \mathcal{V}_1, \ldots, v_n : \mathcal{V}_n\}$ amb restricció que els valors relació que hi pertanyen només poden tenir un tuple $\forall x \in V : |x| = 1$.

La forma canònica s'utilitza per a generalitzar les sèries temporals multivaluades en les operacions on el valor multivaluat no és rellevant. En altres operacions, per exemple la selecció o la junció, el multivalor és rellevant per treballar-hi o per a retornar un resultat on la sèrie temporal és multivaluada. En les sèries temporals multivaluades cada atribut pot ser de tipus diferent, tot i que, com en les sèries temporals canòniques, cada atribut de valor és homogeni.

Hi ha una forma no habitual de les sèries temporals que ocorre quan tenen dos atributs de temps i que anomenem forma doble.

Definició 4.11 (Sèrie temporal doble)**.** *Anomenem sèrie temporal doble a una sèrie temporal que té dos atributs de temps i dos atributs de valors. Sigui* $\{m_0, \ldots, m_k\}$ *una sèrie temporal és doble si cada mesura* m_i *és un tuple* $m_i = (t_1, v_1, t_2, v_2)$ *on* t_1 *i* t_2 *són instants de temps i* v_1 *i* v_2 *són valors. De la mateixa manera, a aquesta mesura* m_i *l'anomenem mesura doble. Sigui* S *una sèrie temporal doble, no té dues parelles de temps repetides* $|\{(t_1, t_2)|(t_1, v_1, t_2, v_2) \in S\}| = |S|$.

La sèrie temporal doble s'utilitza com a càlcul intermedi d'altres operacions com per exemple la junció o el mapatge. En la forma de relació una sèrie temporal doble buida es pot escriure com $\emptyset = (\{t_1 : \mathcal{T}_1, v_1 : \mathcal{V}_1, t_2 : \mathcal{T}_2, v_2 : \mathcal{V}_2, \{\})$.

4.1.5. Exemples

Exemple 4.1 (Valors reals)**.** Sèrie temporal S_1 on el temps i els valors pertanyen a $\bar{\mathbb{R}}$. Conté la mesura de valor 1 en el temps 2, la mesura de valor 3 en el temps 2 i la mesura de valor 1 en el temps 6.

En la forma canònica completa s'escriu com $S_1 = (\{t : \bar{\mathbb{R}}, v : \bar{\mathbb{R}}\}, \{\{(t, 2), (v, 1)\}, \{(t, 3), (v, 3)\}, \{(t, 6), (v, 1)\}\})$. També es pot escriure de manera simplificada com a $S_1 = \{(2, 1), (3, 3), (6, 1)\}$.

La sèrie temporal S_1 es visualitza com a taula a la figura 4.2, a la qual hi afegim una visualització com diagrama de dispersió amb el temps a l'eix horitzontal i el valor a l'eix vertical.

Exemple 4.2 (Valors caràcters)**.** Sèrie temporal S_2 on el temps pertany a $\bar{\mathbb{R}}$ i els valors són caràcters que pertanyen a $C = \{a, b, \ldots, z, \infty\}$. Conté el caràcters a, c i a mesurats respectivament en els temps 2, 3 i 6.

S_1

t	v
2	1
3	3
6	1

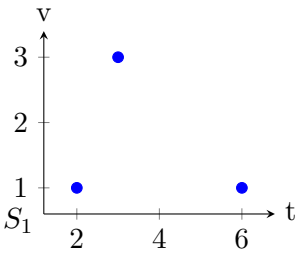


Figura 4.2.: Taula i gràfic d'una sèrie temporal amb valors reals

S_4

t	v
2	a
3	c
6	a

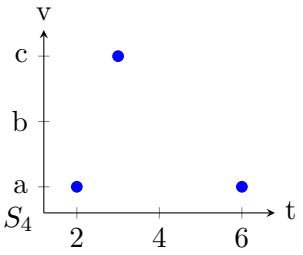


Figura 4.3.: Taula i gràfic d'una sèrie temporal amb valors caràcters

De manera simplificada s'escriu com $S_2 = \{(2, a), (3, c), (6, a)\}$. La sèrie temporal S_2 es visualitza com a taula a la figura 4.3, a la qual hi afegim una visualització com diagrama de dispersió amb el temps a l'eix horitzontal i el valor a l'eix vertical no continu.

Exemple 4.3 (Sèrie temporal multivaluada)**.** Sèrie temporal S_3 on el temps pertany a $\bar{\mathbb{R}}$ i hi ha tres valors on cadascun pertany a $\bar{\mathbb{R}}$. En els temps 2, 3 i 6 s'ha mesurat: a) un atribut *temp* amb valors 1, 2 i 1; b) un atribut *cons* amb valors 2, 1 i 2; i c) un atribut *vol* amb valors 3, 0 i 3.

En la forma multivaluada s'escriu com

$$
\begin{aligned}
S_3 = (\{&t : \bar{\mathbb{R}},\ \text{temp} : \bar{\mathbb{R}},\ \text{cons} : \bar{\mathbb{R}},\ \text{vol} : \bar{\mathbb{R}}\}, \{\\
&\{(t, 2), (\ \text{temp}, 1), (\ \text{cons}, 2), (\ \text{vol}, 3)\},\\
&\{(t, 3), (\ \text{temp}, 2), (\ \text{cons}, 1), (\ \text{vol}, 0)\},\\
&\{(t, 6), (\ \text{temp}, 1), (\ \text{cons}, 2), (\ \text{vol}, 3)\}\\
\})&
\end{aligned}
$$

També es pot escriure de manera simplificada com a $S_3 = ((t,\ \text{temp},\ \text{cons},\ \text{vol}), \{(2, 1, 2, 3), (3, 2, 1, 0), (6, 1, 2, 3)\})$.

S_3

t	temp	cons	vol
2	1	2	3
3	2	1	0
6	1	2	3

S_3^c

t	v		
	temp	cons	vol
2	1	2	3
3	2	1	0
6	1	2	3

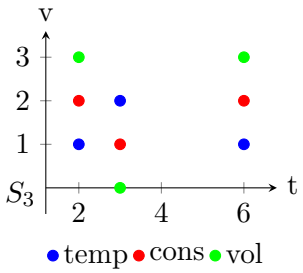


Figura 4.4.: Taula d'una sèrie temporal multivaluada

S_4

t	v
2	(1,2,3)
4	(3,4,5)
6	(1,2,3)

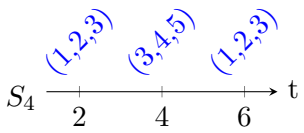


Figura 4.5.: Taula d'una sèrie temporal amb valors vectors

La forma canònica és una sèrie temporal amb tuples (t, v), és a dir

$$
\begin{aligned}
S_3^C = (\{t : \bar{\mathbb{R}}, v : \{\ &\text{temps} : \bar{\mathbb{R}},\ \text{cons} : \bar{\mathbb{R}},\ \text{vol} : \bar{\mathbb{R}},\}\}, \{ \\
&\{(t,2), (v, (\{\ \text{temp} : \bar{\mathbb{R}},\ \text{cons} : \bar{\mathbb{R}},\ \text{vol} : \bar{\mathbb{R}},\}, \{(\ \text{temp}, 1), (\ \text{cons}, 2), (\ \text{vol}, 3)\}\})), \\
&\{(t,3), (v, (\{\ \text{temp} : \bar{\mathbb{R}},\ \text{cons} : \bar{\mathbb{R}},\ \text{vol} : \bar{\mathbb{R}},\}, \{(\ \text{temp}, 2), (\ \text{cons}, 1), (\ \text{vol}, 0)\}\})), \\
&\{(t,6), (v, (\{\ \text{temp} : \bar{\mathbb{R}},\ \text{cons} : \bar{\mathbb{R}},\ \text{vol} : \bar{\mathbb{R}},\}, \{(\ \text{temp}, 1), (\ \text{cons}, 2), (\ \text{vol}, 3)\}\})) \\
\})
\end{aligned}
$$

La sèrie temporal S_3 i la seva forma canònica es visualitzen com a taula a la figura 4.3, a la qual hi afegim una visualització com diagrama de dispersió amb el temps a l'eix horitzontal i els valor a l'eix vertical cadascun amb color diferent.

Exemple 4.4 (Valors vectors)**.** Sèrie temporal S_4 on el temps pertany a $\bar{\mathbb{R}}$ i el valor pertany a $\bar{\mathbb{R}}^3$; és a dir és un vector representat amb un tuple. Conté el valor $(1, 2, 3)$ en el temps 2, el valor $(3, 4, 5)$ en el temps 4 i el valor $(1, 2, 3)$ en el temps 6.

De manera simplificada s'escriu com $S_4 = \{(2, (1, 2, 3)), (3, (3, 4, 5)), (6, (1, 2, 3))\}$ i es visualitza com a taula i com a gràfic a la figura 4.5.

S'observa que una sèrie temporal amb valors vectors és diferent d'una sèrie temporal multivaluada. El domini de la primera són vectors i el de la segona són relacions d'un sol tuple en els que es pot operar cada atribut per separat. En els vectors de forma general no es poden operar cada component per separat sinó que formen una unitat semàntica. Aquesta diferència de significat prové de si es considera que es

S_5

t	v
1	t | v 2 | 1 3 | 3 6 | 1
2	t | v 2 | 1 3 | 3 6 | 1

Figura 4.6.: Taula d'una sèrie temporal amb valors sèrie temporal

mesuren vectors o atributs diferents, el qual s'observa en la visualització: el gràfic d'un vector és un espai R^n en canvi el gràfic d'una sèrie temporal multivaluada és un multigràfic, un gràfic per a cada atribut.

Exemple 4.5 (Valors sèrie temporal)**.** Sèrie temporal S_5 on el temps pertany a $\bar{\mathbb{R}}$ i el valor és una sèrie temporal del mateix format que en l'exemple 1. Conté els tuples de S_1 com a valors en el temps 1 i 2.

De manera simplificada s'escriu com $S_5 = \{(1, \{(2,1),(3,3),(6,1)\}), (2, \{(2,1),\ (3,\ 3),\ (6,1)\})\}$ i es visualitza com a taula a la figura 4.6.

S'observa que la capçalera de S_5 és $\{t : \bar{\mathbb{R}}, v : \{t : \bar{\mathbb{R}}, v : \bar{\mathbb{R}}\}\}$. És a dir, el valor és una altra relació, com es descriu per Date [23, sec. 6.4], on el temps i el valor pertanyen a $\bar{\mathbb{R}}$. Per tant, el valor de S_5 és de tipus sèrie temporal amb valors reals. Això no obstant, el significat d'aquestes sèries temporals és difícil d'interpretar, tot i que són interessants com a exercici acadèmic de definir sèries temporals de sèries temporals.

4.2. Model d'operacions

En aquesta secció definim les operacions d'un SGST que permeten manipular les sèries temporals. Una sèrie temporal té un atribut de temps que ha de ser tingut en compte pels operadors que la manipulin. Així, atenent a aquest atribut de temps, el comportament d'una sèrie temporal pot tenir naturaleses diferents:

- Conjunt, és a dir els operadors només atenent a la forma estructural bàsica.
- Seqüència, en la qual els operadors la tracten com a conjunts amb ordre.
- Funció temporal, en la qual els operadors assumeixen que una sèrie temporal és la representació d'una funció temporal.

En el disseny del model d'operacions següent es distingeix el comportament per als tres casos anteriors. Es dissenyen les operacions bàsiques que permeten que posteriorment es combinin per a elaborar-ne de més complexes.

Les manipulacions de les sèries temporals es defineixen abstractament per a qualsevol sèrie temporal que tingui l'estructura de SGST. Les definicions dels operadors avaluen els conceptes algebraics i lògics de les dades però no avaluen la semàntica en un context particular, com també ocorre en el model d'operacions del model relacional. És a dir, en cada context particular de manipulació d'una sèrie temporal s'ha de decidir si aquella àlgebra té significat o, al contrari, no pot ser aplicada. Per exemple una suma de valors de diferents unitats podria ser semànticament errònia. A la secció 4.3 estudiem el significat d'algunes propietats de les sèries temporals.

4.2.1. Bàsiques de conjunts

En el model estructural d'SGST hem definit les sèries temporals utilitzant conjunts. En aquest apartat definim operadors per a les sèries temporals recollint els operadors habituals que tenen els conjunts.

El model relacional d'SGBDR defineix els seus operadors bàsics a partir de l'àlgebra de conjunts [23, cap. 7]. En aquest apartat apliquem el mateix estudi per al model d'SGST. Tot i així de manera simplificada, a les definicions no es descriuen les sèries temporals com a relacions amb capçaleres sinó que se n'escriuen només els conjunts de valors. Seguint el model relacional es poden estendre les definicions i introduir el model complet de relacions.

En les operacions binàries de sèries temporals, entre les mesures d'ambdós conjunts es poden aplicar les dues relacions d'ordre de la definició 4.4, és a dir poden tenir ordre semitemporal o temporal. Com a conseqüència, aquests dos ordres indueixen dues definicions per a alguns operadors de conjunts. Als operadors amb ordre

temporal els anomenarem temporals i hi afegirem un superíndex t per a indicar-ho.

Les operacions que agrupen els elements dels conjunts habitualment se solen representar gràficament mitjançant diagrames de Venn. A la figura 4.7 es mostren els diagrames de Venn per a cinc de les operacions dels SGSTM que descrivim a continuació: inclusió, unió, diferència, intersecció i diferència simètrica; tant en la seva vessant parcial com en la seva vessant temporal.

Per a dibuixar aquest diagrames de Venn, a banda del conjunt corresponent a cada sèrie temporal S_1 i S_2 i la seva intersecció, cal dibuixar uns subconjunts T_1 i T_2 que indiquen les mesures que comparteixen el mateix temps amb una altra mesura de l'altre conjunt però no el mateix valor. És a dir $T_1 = \{m | m \in S_1 \wedge (\exists n \in S_2 : m =^t n)\}$ i $T_2 = \{m | m \in S_2 \wedge (\exists n \in S_1 : m =^t n)\}$. Aquests dos subconjunts T_1 i T_2 són importants per a les operacions dels SGST perquè no hi pot haver cap sèrie temporal resultant que els inclogui a tots dos, car significaria que conté temps repetits. Per exemple, una operació que tingui la sèrie resultant $T_1 \cup T_2$, com es mostra a la figura 4.8, és impossible

Pertinença i inclusió

La pertinença determina si un element pertany a un conjunt. Sigui S una sèrie temporal i m una mesura, es defineix la pertinença de m a S de la forma habitual en els conjunts. Aquesta pertinença es defineix a partir de l'ordre semitemporal, és a dir que dues mesures són iguals quan ho són els seus temps i valor.

Definició 4.12 (Pertinença)**.** *Sigui S una sèrie temporal i m una mesura, direm que la mesura pertany a la sèrie temporal* $m \in S \iff \exists n \in S : T(m) = T(n) \wedge V(m) = V(n)$.

A partir de l'ordre temporal, és a dir que dues mesures són iguals quan ho són els seus temps, es defineix la pertinença temporal d'una mesura a una sèrie temporal.

Definició 4.13 (Pertinença temporal)**.** *Sigui S una sèrie temporal i m una mesura, direm que la mesura pertany temporalment a la sèrie temporal* $m \in^t S \iff \exists n \in S : T(m) = T(n)$.

Si una mesura pertany a una sèrie temporal, $m \in S$, aleshores també hi pertany temporalment, $m \in^t S$.

La inclusió determina si tots els elements d'un conjunt pertanyen a un altre conjunt. Atenent a la pertinença, es defineix la inclusió d'una mesura a una sèrie temporal.

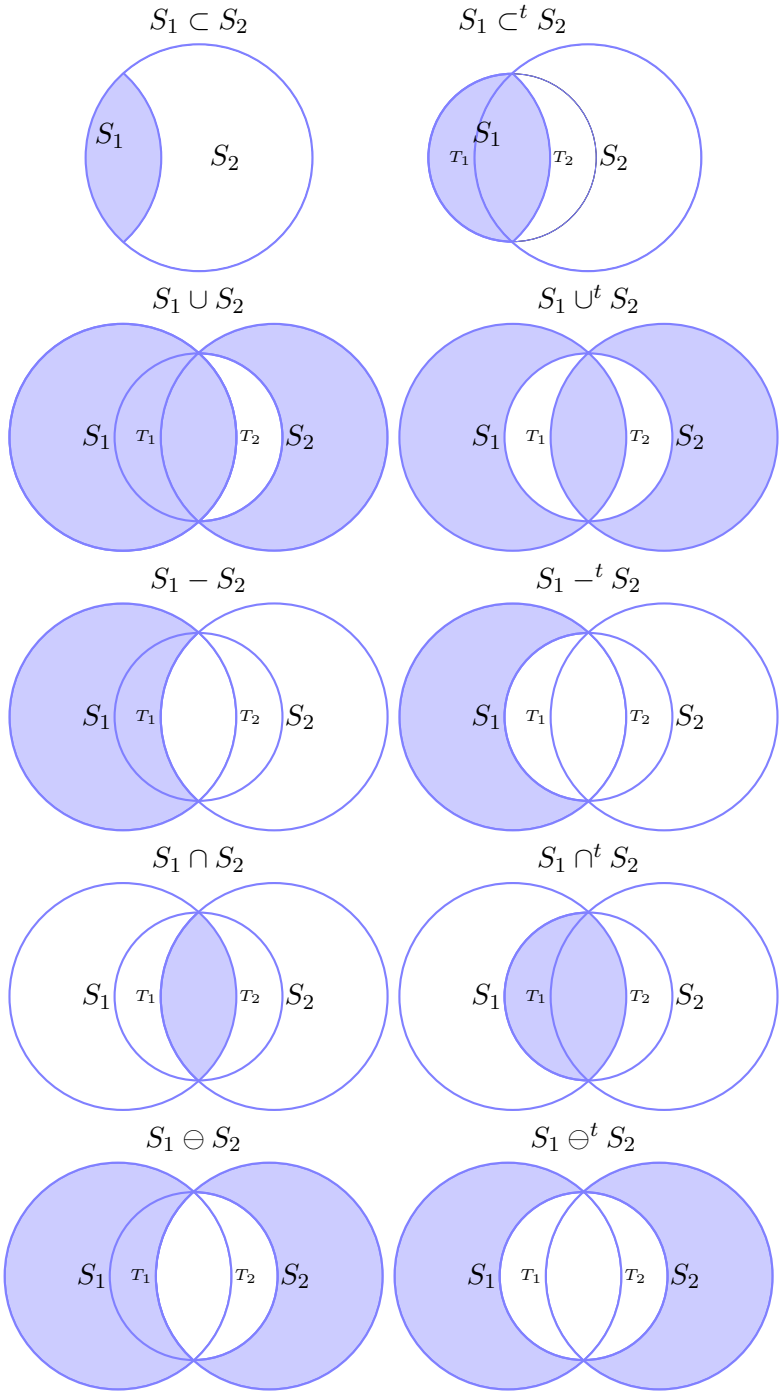


Figura 4.7.: Diagrames de Venn per a les operacions dels SGSTM. Els subconjunts T_1 i T_2 indiquen les mesures $m =^t n$.

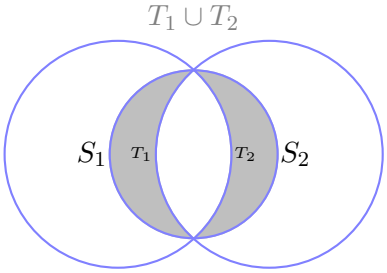


Figura 4.8.: Diagrama de Venn impossible per a les operacions dels SGSTM. Els subconjunts T_1 i T_2 indiquen les mesures $m =^t n$.

Definició 4.14 (Inclusió)**.** *Siguin S_1 i S_2 dues sèries temporals, la primera sèrie temporal està inclosa en la segona $S_1 \subseteq S_2 \iff \forall m \in S_1 : m \in S_2$. Aleshores, S_1 és una subsèrie temporal de S_2.*

Atenent a la pertinença temporal, es defineix la inclusió temporal d'una mesura a una sèrie temporal.

Definició 4.15 (Inclusió temporal)**.** *Siguin S_1 i S_2 dues sèries temporals, la primera sèrie temporal està inclosa temporalment en la segona $S_1 \subseteq^t S_2 \iff \forall m \in S_1 : m \in^t S_2$.*

Exemple 4.6 (Pertinença i inclusió)**.** Siguin les mesures $m_1 = (1, 1)$, $m_2 = (3, 1)$, $m_3 = (3, 2)$ i $m_4 = (4, 0)$ i les sèries temporals $S_1 = \{m_1, m_2\}$ i $S_2 = \{m_1, m_3, m_4\}$ aleshores les operacions següents són certes: $m_2 \in S_1$, $m_2 \notin S_2$, $m_2 \in^t S_1$, $m_2 \in^t S_2$, $S_1 \not\subseteq S_2$ i $S_1 \subseteq^t S_2$.

Màxim i suprem

En una sèrie temporal les mesures tenen relació d'ordre total. Com que la sèrie temporal s'ha considerat finita i sense elements repetits, quan la sèrie temporal no és buida això comporta l'existència d'un màxim i d'un mínim.

Definició 4.16 (Màxim i mínim)**.** *Sigui S una sèrie temporal. El* màxim *de S, notat com a $\max(S)$, és un element de S tal que $\forall m \in S : \max(S) \geq m$. El* mínim *de S, notat com a $\min(S)$, és un element de S tal que $\forall m \in S : \min(S) \leq m$.*

El $\max(S)$ i el $\min(S)$ no estan definits quan la sèrie temporal és buida: $S = \emptyset$. En canvi, com que el domini de temps un conjunt tancat, el suprem i l'ínfim estan definits per qualsevol sèrie temporal. De fet, es poden definir de manera similar al conjunt estès de nombres reals on $\sup(\emptyset) = -\infty$ i $\inf(\emptyset) = +\infty$ [13].

Definició 4.17 (Suprem i ínfim)**.** *Sigui S una sèrie temporal, $m = (-\infty, \infty)$ una mesura indefinida negativa i $n = (+\infty, \infty)$ una mesura indefinida positiva.*

El suprem *de S, notat com a* $\sup(S)$*, és*

$$\sup(S) = \begin{cases} m & \textit{quan } S = \emptyset \\ \max(S) & \textit{altrament} \end{cases}$$

*L'*ínfim *de S, notat com a* $\inf(S)$*, és*

$$\inf(S) = \begin{cases} n & \textit{quan } S = \emptyset \\ \min(S) & \textit{altrament} \end{cases}$$

Quan la sèrie temporal no és buida, per ser un conjunt finit i d'ordre total, sempre hi ha un i només un màxim i un mínim i per tant es corresponen amb el suprem i l'ínfim respectivament.

Exemple 4.7 (Mínim i suprem)**.** Siguin les sèries temporals $S_1 = \{(1, 1), m_2 = (3, 1)\}$ i $S_2 = \{\}$ aleshores les operacions següents són certes: $\min(S_1) = \inf(S_1) = (1, 1)$ i $\sup(S_2) = (-\infty, \infty)$.

Unió

La unió de dos conjunts és un conjunt que conté tots els elements d'ambdós conjunts. Per a poder unir dos conjunts amb estructura de relació, $A \cup B$, cal que tots dos tinguin la mateixa estructura; és a dir, en termes de SGBDR cal que A i B tinguin la mateixa capçalera.

Per tal que l'operació d'unió de conjunts sigui vàlida per a les sèries temporals cal, a més, tenir en compte quan dues sèries temporals tenen mesures en el mateix instant de temps. En cas d'utilitzar l'operació d'unió de conjunts la sèrie temporal resultant no compliria amb la definició 4.7 ja que contindria mesures amb temps repetits. Com a conseqüència, es defineixen dues operacions d'unió per a les sèries temporals que resolen la restricció del temps de forma diferent. Per a definir ambdues unions cal usar la pertinença temporal ja que cal treballar amb els conjunts que comparteixen instants de temps, per tant és difícil establir la referència de pertinença per a cada una. Definim la primera unió com la més propera possible a la unió de conjunts.

En primer lloc, es defineix la unió de dues sèries temporals que escull les mesures del primer operand en cas de mesures amb el mateix temps però diferent valor.

Definició 4.18 (Unió)**.** *Siguin S_1 i S_2 dues sèries temporals en què* $\operatorname{dom} S_1 = \operatorname{dom} S_2$*, la unió de les dues sèries temporals, notada com a $S_1 \cup S_2$, és una sèrie temporal que conté totes les mesures de S_1 i les mesures de S_2 que no tenen temps repetits:* $S_1 \cup S_2 = \{m | m \in S_1 \vee (m \in S_2 \wedge m \notin^t S_1)\}$.

Propietats de la unió de sèries temporals:

- El cardinal de la sèrie temporal resultant està fitat a $|S_1| \leq |S| \leq |S_1| + |S_2|$.
- No commutativa. En general $S_1 \cup S_2 \neq S_2 \cup S_1$ tot i que sí que es compleix l'equivalència respecte al cardinal $|S_1 \cup S_2| = |S_2 \cup S_1|$.

En segon lloc, es defineix la unió temporal de dues sèries temporals, la qual és la unió sense tenir en compte les mesures que tenen el mateix instant de temps i diferent valor.

Definició 4.19 (Unió temporal)**.** *Siguin S_1 i S_2 dues sèries temporals en què* $\operatorname{dom} S_1 = \operatorname{dom} S_2$*, la unió temporal de les dues sèries temporals, notada com a $S_1 \cup^t S_2$, és una sèrie temporal que conté les mesures de S_1 i de S_2 excloent les que només comparteixen el temps: $S_1 \cup^t S_2 = \{m | (m \in S_1 \wedge m \in S_2) \vee (m \in S_2 \wedge m \notin^t S_2) \vee (m \in S_2 \wedge m \notin^t S_1)\}$.*

Propietats de la unió temporal:

- Commutativa

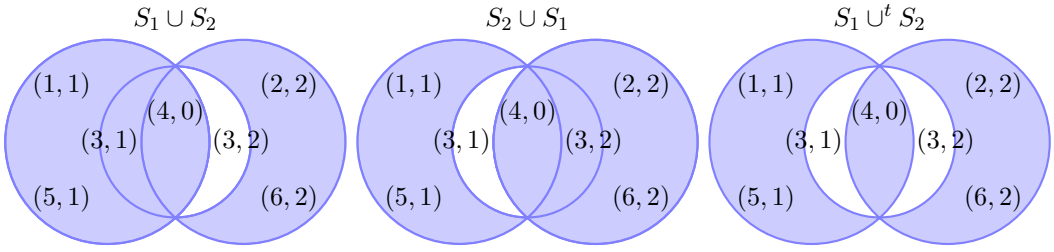
Exemple 4.8 (Unió de dues sèries temporals)**.** Siguin les dues sèries temporals $S_1 = \{(1,1),(3,1),(4,0),(5,1)\}$ i $S_2 = \{(2,2),(3,2),(4,0),(6,2)\}$. La unió de la primera amb la segona és $S_1 \cup S_2 = \{(1,1),(2,2),(3,1),(4,0),(5,1),(6,2)\}$ i, com que no és commutativa, la unió de segona amb la primera és $S_2 \cup S_1 = \{(1,1),(2,2),(3,2),(4,0),(5,1),(6,2)\}$. La unió temporal de totes dues, que és commutativa, és $S_1 \cup^t S_2 = S_2 \cup^t S_1 = \{(1,1),(2,2),(4,0),(5,1),(6,2)\}$.

A la figura 4.9 es mostren els diagrames Venn per a les tres operacions, on l'àrea pintada és la sèrie temporal resultant. L'àrea central d'intersecció dels dos conjunts són les mesures que comparteixen temps i valor, en aquest cas la mesura $(4,0)$. L'àrea central esquerra són les mesures de S_1 que només comparteixen temps amb una mesura de S_2, és a dir la $(3,1)$, i dualment a l'àrea central dreta hi ha la $(3,2)$. Les àrees més externes es corresponen amb la resta de mesures. A la figura 4.9 també es mostren les mateixes operacions amb la visualització en taula de les sèries temporals.

Diferència

La diferència de dos conjunts és un conjunt que conté tots els elements del primer conjunt que no pertanyen al segon. Per a poder restar dos conjunts amb estructura de relació, $A - B$, cal que tots dos tinguin la mateixa estructura; és a dir, en termes de SGBDR cal que A i B tinguin la mateixa capçalera. En la definició de l'operació de diferència cal tenir en compte les dues pertinences possibles.

En primer lloc, es defineix la diferència atenent a la pertinença estricta de conjunts. És a dir s'aplica la diferència de conjunts a les sèries temporals.



S_1

t	v
1	1
3	1
4	0
5	1

S_2

t	v
2	2
3	2
4	0
6	2

$S_1 \cup S_2$

t	v
1	1
2	2
3	1
4	0
5	1
6	2

$S_2 \cup S_1$

t	v
1	1
2	2
3	2
4	0
5	1
6	2

$S_1 \cup^t S_2$

t	v
1	1
2	2
4	0
5	1
6	2

Figura 4.9.: Diagrames Venn i taules per als exemples d'unió i d'unió temporal

Definició 4.20 (Diferència)**.** *Siguin S_1 i S_2 dues sèries temporals en què* $\operatorname{dom} S_1 = \operatorname{dom} S_2$*, la diferència de les dues sèries temporals, $S_1 - S_2$, és una sèrie temporal que conté totes les mesures de S_1 que no pertanyen a S_2: $S_1 - S_2 = \{m | m \in S_1 \wedge m \notin S_2\}$.*

En segon lloc, es defineix la diferència atenent a la pertinença temporal.

Definició 4.21 (Diferència temporal)**.** *Siguin S_1 i S_2 dues sèries temporals en què* $\operatorname{dom} S_1 = \operatorname{dom} S_2$*, la diferència temporal de les dues sèries temporals, $S_1 -^t S_2$, és una sèrie temporal que conté totes les mesures de S_1 que no pertanyen temporalment a S_2: $S_1 -^t S_2 = \{m | m \in^t S_1 \wedge m \notin^t S_2\}$.*

Intersecció

La intersecció de dos conjunts és un conjunt que conté els elements comuns als dos conjunts. Per a poder intersecar dos conjunts amb estructura de relació, $A \cap B$, cal que tots dos tinguin la mateixa estructura; és a dir, en termes de SGBDR cal que A i B tinguin la mateixa capçalera.

En la definició de l'operació d'intersecció cal tenir en compte les dues pertinences possibles.

En primer lloc, es defineix la diferència atenent a la pertinença estricta de conjunts. És a dir s'aplica l'operació d'intersecció de conjunts.

Definició 4.22 (Intersecció)**.** *Siguin S_1 i S_2 dues sèries temporals en què* $\operatorname{dom} S_1 = \operatorname{dom} S_2$*, la intersecció de les dues sèries temporals, $S_1 \cap S_2$, és una sèrie temporal que conté les mesures de S_1 repetides a S_2: $S_1 \cap S_2 = \{m | m \in S_1 \wedge m \in S_2\}$.*

En segon lloc, es defineix la intersecció atenent a la pertinença temporal tenint en compte quan dues sèries temporals tenen mesures en el mateix instant de temps però de valor diferent.

Definició 4.23 (Intersecció temporal)**.** *Siguin S_1 i S_2 dues sèries temporals en què* $\operatorname{dom} S_1 = \operatorname{dom} S_2$*, la intersecció temporal de les dues sèries temporals, $S_1 \cap^t S_2$, és una sèrie temporal que conté les mesures de S_1 repetides temporalment a S_2: $S_1 \cap^t S_2 = \{m | m \in^t S_1 \wedge m \in^t S_2\}$.*

Propietats de la intersecció:

- La intersecció és commutativa però la intersecció temporal no és commutativa.
- A partir de la diferència es pot definir la intersecció: $S_1 \cap S_2 = S_1 - (S_1 - S_2)$.

Diferència simètrica

La diferència simètrica de dos conjunts és un conjunt que conté els elements no comuns dels dos conjunts. La diferència simètrica de dos conjunts $A \ominus B$ es defineix a partir de la diferència i la unió:

$$\begin{aligned} A \ominus B &= (A - B) \cup (B - A) \\ &= (A \cup B) - (A \cap B) \\ A \ominus B &\subseteq A \cup B \end{aligned}$$

Seguint aquestes propietats es defineixen dues diferències simètriques: una a partir de la diferència i la unió de sèries temporals i una altra a partir de la diferència temporal i la unió temporal. Per tal que l'operació de diferència simètrica sigui vàlida per a les sèries temporals cal tenir en compte quan dues sèries temporals tenen mesures en el mateix instant de temps.

En primer lloc, es defineix la diferència simètrica excloent les mesures amb el mateix temps però de valor diferent.

Definició 4.24 (Diferència simètrica)**.** *Siguin S_1 i S_2 dues sèries temporals en què* $\operatorname{dom} S_1 = \operatorname{dom} S_2$*, la diferència simètrica de les dues sèries temporals, $S_1 \ominus S_2$, és una sèrie temporal que conté les mesures de S_1 o exclusivament les de S_2: $S_1 \ominus S_2 = \{m | (m \in S_1 \wedge m \notin S_2) \vee (m \in S_2 \wedge m \notin^t S_1)\}$.*

En segon lloc, es defineix la diferència simètrica temporal excloent les mesures amb el mateix temps.

Definició 4.25 (Diferència simètrica temporal)**.** *Siguin S_1 i S_2 dues sèries temporals en què* $\operatorname{dom} S_1 = \operatorname{dom} S_2$*, la diferència simètrica de les dues sèries temporals, $S_1 \ominus^t S_2$, és una sèrie temporal que conté les mesures de S_1 o exclusivament les de S_2: $S_1 \ominus^t S_2 = \{m | (m \in^t S_1 \wedge m \notin^t S_2) \vee (m \in^t S_2 \wedge m \notin^t S_1)\}$.*

Selecció

La selecció és una operació dels SGBDR que selecciona uns tuples determinats d'un conjunt, a vegades també s'anomena restricció.

Definició 4.26 (Selecció)**.** *Sigui la sèrie temporal S, a_1 i a_2 dos noms d'atributs que pertanyen a S, i $a_1 \Theta a_2$ una expressió booleana sobre a_1 i a_2, la selecció de S per l'expressió booleana s'escriu com $\sigma_{a_1 \Theta a_2}(S)$ i es defineix de la mateixa manera que en els* SGBDR *[23, cap. 7].*

En una forma més genèrica, l'expressió booleana pot incloure un o més atributs i està formada per més d'una expressió lògica.

Exemple 4.9 (Selecció de les mesures majors a un instant de temps)**.** Sigui la sèrie temporal $S_1 = \{(1,1),(3,1),(4,0),(5,1)\}$, la selecció dels temps més grans que 3 és $\sigma_{t>3}(S_1) = \{(4,0),(5,1)\}$.

Projecció

La projecció és una operació dels SGBDR que selecciona uns atributs determinats d'un conjunt. Aquesta operació treballa amb la capçalera de la sèrie temporal, és a dir amb els atributs que genèricament són t i v però que també poden tenir altres noms.

Definició 4.27 (Projecció)**.** *Sigui la sèrie temporal S i sigui $A = \{a_0, \ldots, a_n\}$ un conjunt de noms d'atributs, la projecció de la sèrie temporal en els atributs s'escriu com $\Pi_A(S)$ i es defineix de la mateixa manera que en els* SGBDR *[23, cap. 7]. Aleshores aquesta nova sèrie temporal $\Pi_A(S)$ només inclou els atributs A de les mesures.*

En l'operació de projecció, si els atributs seleccionats no inclouen l'atribut temps o només inclouen un atribut el resultat no és una sèrie temporal sinó que és un conjunt relacional.

Exemple 4.10 (Projecció d'alguns atributs de la sèrie temporal)**.** Sigui la sèrie temporal $S_1 = \{(1,1),(3,1),(4,0),(5,1)\}$, la projecció en l'atribut de temps és el conjunt $\Pi_{\{t\}}(S_1) = \{1,3,4,5\}$. Sigui la sèrie temporal multivaluada $S_2 = ((t,\ \text{temp},\ \text{cons},\ \text{vol}), \{(2,1,2,3),(3,2,1,0),(6,1,2,3)\})$, la projecció en els atributs t i *temp* és la sèrie temporal $\Pi_{\{t,\text{temp}\}}(S_1) = ((t,\ \text{temp}), \{(2,1),(3,2),(6,1)\})$

Reanomena

El reanomena és una operació dels SGBDR que canvia el nom dels atributs. Aquesta operació treballa amb la capçalera de la sèrie temporal.

Definició 4.28 (Reanomena)**.** *Sigui la sèrie temporal S, a un nom d'atribut que pertany a S i b un que no hi pertany, reanomenar a per b s'escriu com $\rho_{a/b}(S)$ i es defineix de la mateixa manera que en els* SGBDR *[23, cap. 7].*

En una forma més genèrica, es poden reanomenar més d'un atribut alhora.

Exemple 4.11 (Reanomena els atributs de la sèrie temporal)**.** Sigui la sèrie temporal multivaluada $S_2 = ((t,\ \text{temp},\ \text{cons},\ \text{vol}), \{(2,1,2,3),(3,2,1,0),(6,1,2,3)\})$, reanomenar l'atribut *temp* per *v1* és la sèrie temporal $\rho_{\text{temp/v1}}(S_1) = ((t,\ \text{v1},\ \text{cons},\ \text{vol}), \{(2,1,2,3),(3,2,1,0),(6,1,2,3)\})$.

Producte i junció

El producte cartesià de dos conjunts és un conjunt que conté totes les parelles possibles d'elements d'ambdós conjunts. Per a poder multiplicar dos conjunts amb estructura de relació, $A \times B$, en termes de SGBDR cal que A i B no tinguin en comú noms d'atributs. En els SGBDR, a diferència del producte de conjunts, el conjunt resultant no és un conjunt de parells de tuples sinó un conjunt de tuples.

Definim el producte de dues sèries temporals, les qual en forma canònica tinguin els atributs t i v, com una sèrie temporal amb atributs t_1, v_1, t_2 i v_2. Així doncs, per a sèries temporals el producte resulta en una sèrie temporal amb dos atributs de temps, a la qual anomenem sèrie temporal doble (v. definició 4.11).

Definició 4.29 (Producte)**.** *Siguin S_1 i S_2 dues sèries temporals en forma canònica, el producte de les dues sèries temporals $S_1 \times S_2$ és una sèrie temporal doble que conté la unió de totes les parelles de mesures de S_1 i S_2: $S_1 \times S_2 = \{(t_1, v_1, t_2, v_2) | (t_1, v_1) \in S_1 \wedge (t_2, v_2) \in S_2\}$*

Propietats del producte:

- El cardinal resultant és $|S| = |S_1||S_2|$
- El grau resultant és 4

La junció (*join*) de dos conjunts és un conjunt que conté les parelles d'elements d'ambdós conjunts que tenen el mateix valor per als atributs comuns. La junció de dos conjunts amb estructura de relació es defineix com una selecció sobre el producte [23, cap. 7].

Per a les sèries temporals, definim la junció com l'ajuntament de les parelles que tenen el mateix atribut de temps en ambdues sèries temporals . El resultat de la junció és una sèrie temporal multivaluada.

Definició 4.30 (Junció)**.** *Siguin S_1 i S_2 dues sèries temporals en forma canònica, la junció de les dues sèries temporals, $S_1 \bowtie S_2$, és una sèrie temporal multivaluada que selecciona del producte de S_1 amb S_2 les mesures dobles amb temps iguals:* $S_1 \bowtie S_2 = \{(t, v_1, v_2) | (t, v_1) \in S_1 \wedge (t, v_2) \in S_2\}$.

Propietats de la junció:

- $\mathrm{dom}(S_1 \bowtie S_2) = \mathrm{dom}\, S_1 \times \mathrm{dom}\, S_2$
- El cardinal resultant és $|S| \leq \min(|S_1|, |S_2|)$
- És commutativa; tenint en compte que els atributs tenen nom i per tant l'ordre no importa.

Cal tenir en compte que la junció només sap operar amb dues sèries temporals que tinguin el mateix vector de temps; és a dir regulars entre elles (v. definició 4.63). En el cas que no tinguin el mateix vector de temps, es pot aplicar la junció temporal de la definició 4.42.

Exemple 4.12 (Junció de dues sèries temporals)**.** Siguin les dues sèries temporals $S_1 = \{(1,1), (3,1), (4,0), (5,1)\}$ and $S_2 = \{(2,2), (3,2), (4,0), (6,2)\}$. La junció de totes dues és $S_1 \bowtie S_2 = \{(3,1,2), (4,0,0)\}$.

Computacionals: mapa, agregació i plec

Per a poder operar amb els conjunts, a més de l'àlgebra definida fins ara, es necessiten operadors amb funcionalitats computacionals; és a dir, operadors que calculin amb els valors continguts en els conjunts.

En els SGBDR els operadors computacionals bàsics són *extend*, *aggregate* i *summarize* [23, cap. 7]. Per a les sèries temporals definim operacions equivalents a les dues primeres de la manera amb què habitualment s'utilitzen per als conjunts. La tercera, el *summarize*, és una operació que s'utilitza per a sintetitzar informació mitjançant grups, és a dir aplica operacions *aggregate* a conjunts que prèviament s'han agrupat segons un atribut compartit. Per a les sèries temporals, però, necessitem una operació computacional més genèrica que ens permeti calcular recursivament sense haver de definir grups.

Així doncs, a continuació es defineix l'operador mapa (*map*) com a equivalent a l'*extend*, l'operador agregació (*aggregate*) com a equivalent a l'*aggregate* i l'operador plec (*fold*) com una forma més general de calcular recursivament amb les mesures que *summarize*.

L'operació de mapatge aplica una funció a cada element del conjunt.

Definició 4.31 (Mapa)**.** *Sigui S una sèrie temporal en què $\mathcal{V} = \operatorname{dom} S$ i sigui $f : \mathcal{T} \times \mathcal{V} \to \mathcal{T}' \times \mathcal{V}'$ una funció sobre una mesura que retorna una mesura. El* mapa *de f a S és una sèrie temporal amb la funció aplicada a cada mesura:* $\text{mapa}(S, f) = \{f(m) | m \in S\}$*. Noteu que* $\operatorname{dom}(\text{mapa}(S, f)) = \mathcal{V}'$*.*

L'operació d'agregació sintetitza en una mesura la informació dels elements del conjunt segons un criteri, per exemple estadístics.

Definició 4.32 (Agregació)**.** *Sigui $S = \{m_0, \ldots, m_k\}$ una sèrie temporal en què $\mathcal{V} = \operatorname{dom} S$, sigui m una mesura amb $\mathcal{V} = \operatorname{dom} m$ i sigui $f : (\mathcal{T} \times \mathcal{V}) \times (\mathcal{T} \times \mathcal{V}) \to \mathcal{T} \times \mathcal{V}$ una funció sobre dues mesures que retorna una mesura. L'*agregació *de S segons f amb valor inicial m és una mesura que sintetitza la informació de les mesures:* $\text{agregació}(S, m, f) = f(\cdots(f(f(f(m, m_0), m_1), m_2)\cdots), m_k)$*.*

L'operació de plegament combina recursivament els elements del conjunt segons un criteri. Cal notar que l'agregació definida anteriorment és un cas específic del plegament. Assumiu que $\mathcal{P}(C)$ és el conjunt potència (*powerset*) de C.

Definició 4.33 (Plec)**.** *Siguin $S = \{m_0, \ldots, m_k\}$ i R dues sèries temporals en les quals $\mathcal{V} = \operatorname{dom} S$ i $\mathcal{V}' = \operatorname{dom} R$, i sigui $f : \mathcal{P}(\mathcal{T} \times \mathcal{V}') \times (\mathcal{T} \times \mathcal{V}) \to \mathcal{P}(\mathcal{T} \times \mathcal{V}')$ una funció sobre una sèrie temporal i una mesura que retorna una sèrie temporal. El* plec *de S per f amb valor inicial R és una sèrie temporal amb les mesures combinades:* $\text{plec}(S, R, f) = f(\cdots(f(f(f(R, m_0), m_1), m_2)\cdots)m_k)$*.*

Les operacions d'agregació i plegament tal com s'han definit es realitzen en ordre aleatori de mesures. Segons el criteri que s'utilitzi, l'ordre és important i per tant cal una operació que computi tenint-lo en compte. A tal efecte, a continuació s'amplia la definició de la funció de plegament per a tenir en compte l'ordre; per a la funció d'agregació es pot aplicar el mateix concepte.

Definició 4.34 (Plec amb ordre)**.** *Siguin $S = \{m_0, \ldots, m_k\}$ i R dues sèries temporals en les quals $\mathcal{V} = \operatorname{dom} S$ i $\mathcal{V}' = \operatorname{dom} R$, sigui $f : \mathcal{P}(\mathcal{T} \times \mathcal{V}') \times (\mathcal{T} \times \mathcal{V}) \to \mathcal{P}(\mathcal{T} \times \mathcal{V}')$ una funció sobre una sèrie temporal i una mesura que retorna una sèrie temporal, i sigui $g : \mathcal{P}(\mathcal{T} \times \mathcal{V}) \to (\mathcal{T} \times \mathcal{V})$ una funció que retorna una mesura d'una sèrie temporal. El* plec *de S per f amb valor inicial R i ordre g és una sèrie temporal que combina les mesures seguint l'ordre:*

$$\text{oplec}(S, R, f, g) = \begin{cases} R & \textit{si } |S| = 0, \\ \text{oplec}(Q, f(R, q), f, g) & \textit{altrament} \end{cases}$$

on $q = g(S)$ i $Q = S - \{q\}$.

El plec amb ordre és necessari quan la funció f no és associativa ni commutativa perquè llavors l'ordre dels càlculs és important.

Propietats de les operacions computacionals:

- El plec sense ordre és un plec amb ordre aleatori: $\text{plec}(S, R, f) = \text{oplec}(S, R, f, \text{aleatori})$.
- El plec d'una sèrie temporal buida és la sèrie inicial: $\text{plec}(\emptyset, R, f) = R$.
- El plec per una funció que sempre retorni la sèrie inicial és la sèrie inicial: $\text{plec}(S, R, f) = R$ on $f(Q, m) = Q$.
- El plec per una funció que només retorni la mesura original és una sèrie amb una sola mesura: $\text{plec}(S, R, f) = S'$ on $f(Q, m) = \{m\}$ i $|S'| = 1$.
- La funció d'unió en el plegament permet fer la identitat: $S = \text{plec}(S, \emptyset, f)$ on $f(Q, m) = Q \cup \{m\}$.
- Els mapes es poden implementar com a plecs: $\text{mapa}(S, f) = \text{plec}(S, \emptyset, g)$ on $g(Q, m) = \{f(m)\} \cup Q$. De manera semblant, Lämmel [76] també exemplifica com els mapes es poden implementar com a plecs.
- Les agregacions es poden implementar com a plecs: $\text{agregació}(S, m, f) = \text{plec}(S, \{m\}, g)$ on $g(\{n\}, o) = \{f(n, o)\}$.

Exemple 4.13 (Mapes de sèries temporals)**.** Definicions de funcions d'exemple a partir de l'operació computacional de mapatge:

- $\text{identitat}(S) = \text{mapa}(S, f)$ on $f(t, v) = (t, v)$
- $\text{intercanvi}(S) = \text{mapa}(S, f)$ on $f(t, v) = (v, t)$
- $\text{duplica\_t}(S) = \text{mapa}(S, f)$ on $f(t, v) = (t, t)$
- $\text{translació}(S, s) = \text{mapa}(S, f)$ on $f(t, v) = (t+s, v)$ i s és una durada de temps
- $\text{multiplica\_tv}(S) = \text{mapa}(S, f)$ on $f(t, v) = (t, t \cdot v)$

Exemple 4.14 (Agregacions de sèries temporals)**.** Definicions de funcions d'exemple a partir de l'operació computacional d'agregació. Ens els exemples següents utilitzem la notació descrita anteriorment $f(m, n)$ per a definir les funcions d'agregació, en què m pot ser la mesura inicial o les mesures resultants i n és una mesura pertanyent a la sèrie temporal agregada.

- $\text{cardinal}(S) = V\big(\text{agregació}(S, (0, 0), f)\big)$ on $f(m, n) = (0, V(m) + 1)$. Aquesta funció és una implementació del cardinal de la definició 4.8 a partir de l'agregació. Noteu que l'atribut de temps no té cap sentit en aquesta computació.
- $\text{suma\_v}(S) = V\big(\text{agregació}(S, (0, 0), f)\big)$ on $f(m, n) = (0, V(m) + V(n))$. Noteu que en aquesta computació l'atribut de temps tampoc té cap sentit.
- $\text{mitjana\_v}(S) = \text{suma\_v}(S) / \text{cardinal}(S)$
- $\sup(S) = \text{agregació}(S, (-\infty, \infty), f)$ on $f(m, n) = m$ si $T(n) < T(m)$ o $f(m, n) = n$ en cas contrari. Aquesta funció és una implementació de l'operació suprem de la definició 4.17 a partir de l'agregació.

- max_v$(S) = V($ agregació$(S, (0, -\infty), f))$ on $f(m, n) = (0, \max(V(m), V(n))$. A diferència del sup$(S)$ o del max(S), el max_v(S) calcula el màxim dels valors.

Exemple 4.15 (Plecs de sèries temporals)**.** Definicions de funcions d'exemple a partir de l'operació computacional de plegament.

- tpredecessors(S) = plec(S, S, f) on $f(R, m) = \{(T(m), s)\} \cup R$ i $s = T(\sup($ $\sigma_{t<T(m)}R))$. Per a cada mesura (t, v) de la sèrie temporal, el resultat conté una mesura (t, s) en què s és el temps de la mesura precedent a (t, v).
- vpredecessors(S) = mapa(tpredecessors$(S), f)$ on $f(m) = (T(m), V(\sup(R))$ i $R = \sigma_{t=V(m)}S$. Per a cada mesura de la sèrie temporal indica quin és el valor de la mesura precedent, és una definició a partir de l'operació de tpredecessors.

Exemple 4.16 (Aplicacions de les operacions computacionals)**.** Sigui la sèrie temporal $S = \{(1, 1), (3, 1), (4, 0), (5, 1)\}$. La duplicació dels temps en els valors de la sèrie temporal és duplica_t$(S) = \{(1, 1), (3, 3), (4, 4), (5, 5)\}$. La mitjana dels valors de la sèrie temporal és mitjana_v$(S) = 0{,}75$. Els temps predecessors de cada mesura de la sèrie temporal són tpredecessors$(S) = \{(1, -\infty), (3, 1), (4, 3), (5, 4)\}$ i els valors predecessors són vpredecessors$(S) = \{(1, \infty), (3, 1), (4, 1), (5, 0)\}$.

Computacionals binàries amb els valors

Una operació en els conjunts és la que aplica un operador binari a totes les parelles possibles dels elements de dos conjunts. Per exemple la suma, és a dir l'operador binari $+$, aplicada a dos conjunts A i B és un conjunt $A + B = \{a + b | (a, b) \in A \times B\}$.

Per a les sèries temporals també calen operacions computacionals amb les mesures de dues sèries temporals. En el cas d'operar amb dues sèries temporals primer cal ajuntar les dues sèries temporals que es volen operar i després aplicar les operacions computacionals binàries a la sèrie temporal resultant.

El producte i la junció són els operadors que permeten crear parelles de mesures de dues sèries temporals. Per a operar amb els valors de dues sèries temporals la junció és més adequada ja que permet ajuntar el valors que tenen temps comuns. Així doncs, es defineix l'aplicació d'un operador binari de valors a dues sèries temporals a partir de la junció.

Definició 4.35 (Operació computacional binària amb els valors)**.** *Siguin S_1 i S_2 dues sèries temporals i sigui $\odot$ un operador binari en el domini dels valors, $\odot : \mathcal{V}_1 \times \mathcal{V}_2 \to \mathcal{V}'$. L'aplicació d'aquest operador binari a dues sèries temporals és $S_1 \odot S_2 =$* mapa$(S_1 \bowtie S_2, f)$ *on $f(t, v, w) = (t, v \odot w)$.*

Exemple 4.17 (Aplicacions de les operacions computacionals binàries)**.** Exemples de l'aplicació d'operacions computacionals binàries en què s'aplica un operador binari $\odot$ als valors de dues sèries temporals

- $S' = S_1 + S_2$
- $S' =$ subtracció(S_1, S_2). Noteu que indiquem amb el nom complet, subtracció, l'operació computacional de dues sèries temporals corresponent a l'operació aritmètica de resta per tal de no confondre-la amb l'operació de diferència de conjunts que té el guionet per símbol (-).

Les operacions computacionals binàries també es poden usar per a definir altres operacions, per exemple per a calcular els increments de valor d'una sèrie temporal increments$(S) =$ subtracció$(S,$ vpredecessors$(S))$.

Exemple 4.18 (Suma de dues sèries temporals i increments d'una)**.** Siguin les dues sèries temporals $S_1 = \{(1,1),(3,1),(4,0),(5,1)\}$ and $S_2 = \{(1,2),(3,2),(4,0),(5,2)\}$. La suma de les dues sèries temporals és $S_1 + S_2 = \{(1,3),(3,3),(4,0),(5,3)\}$. Els increments de la primera sèrie temporal són increments$(S_1) = \{(1,\infty),(3,0),(4,-1),(5,1)\}$.

4.2.2. Bàsiques de seqüències

Atesa la relació d'ordre induïda pel temps en una sèrie temporal (def. 4.4), les sèries temporals es poden tractar com a seqüències. En aquest apartat definim operadors per a les sèries temporals recollint els operadors habituals que tenen les seqüències.

Els operadors que treballen amb seqüències tenen en compte l'atribut que marca un ordre total en el conjunt. En el cas de les sèries temporals aquest atribut és el temps.

Interval

L'interval sobre una seqüència és la subseqüència compresa entre dos elements. Per a les sèries temporals és possible definir el concepte d'interval sobre la seqüència com la subsèrie entre dos instants de temps, semblant a com es fa a [57, 70]. És una operació de selecció però amb la notació habitual en les seqüències.

Definició 4.36 (Interval)**.** *Sigui S una sèrie temporal i siguin s i t dos instants de temps. Definim el subconjunt $S(s,t) \subseteq S$ com la sèrie temporal $S(s,t) = \{m | m \in S \wedge s < T(m) < t\}$.*

Tal com es fa en les seqüències, es defineix una notació de parèntesis i claudàtors per indicar si l'interval és obert, tancat o semiobert:

$S[s,t) = \{m | m \in S \wedge s \leq T(m) < t\}$

$S(s,t] = \{m | m \in S \wedge s < T(m) \leq t\}$

$S[s,t] = \{m | m \in S \wedge s \leq T(m) \leq t\}$

Propietats:

- La subsèrie $S[-\infty,t) \subseteq S$ és equivalent a la sèrie temporal $S[-\infty,t) = S[T(\inf(S)),t)$. De la mateixa manera $S(s,+\infty] = S(s,T(\sup(S))]$.
- L'interval degenerat $S[t,t] \subseteq S$ és equivalent a la sèrie temporal $S[t,t] = \{m | m \in S \wedge T(m) = t\}$. Els intervals $S(t,t] \subseteq S$ i $S[t,t) \subseteq S$ són equivalents a la sèrie temporal buida $S(t,t] = S[t,t) = \emptyset$ ja que per ser els temps d'ordre total $\nexists T(m) : t < T(m) \leq t$ o $\nexists T(m) : t \leq T(m) < t$, respectivament.
- La subsèrie $S[-\infty,+\infty] \subseteq S$ és equivalent a la sèrie temporal original $S[-\infty, +\infty] = S$. La subsèrie $S(-\infty,+\infty) \subseteq S$ només és equivalent a la sèrie temporal original quan aquesta no conté mesures indefinides $S(-\infty,+\infty) \iff S : (-\infty,v) \notin S \wedge (+\infty,w) \notin S$ on v i w són dos valors qualssevol.

Successió

Atenent a la relació d'ordre induïda pel temps en una sèrie temporal, es defineix el concepte de successor i predecessor en una seqüència. A partir d'una mesura, aquests conceptes determinen quina és la mesura immediatament següent i la mesura immediatament anterior contingudes en una sèrie temporal.

Definició 4.37 (Successor i predecessor)**.** *Sigui S una sèrie temporal i sigui m una mesura. El successor de m en S, notat com* $\text{seg}_S(m)$*, és* $\text{seg}_S(m) = \inf(S(T(m), +\infty])$*. El predecessor de m en S, notat com* $\text{ant}_S(m)$*, és* $\text{ant}_S(m) = \sup(S[-\infty, T(m)))$*.*

Quan no hi hagi dubte de la sèrie temporal que marca l'ordre, per exemple quan $m \in S$, podem escriure $\text{seg}(m)$ *i* $\text{ant}(m)$*.*

S'observa que s'obtenen mesures indefinides en els casos que la mesura següent o anterior es calcula respectivament per la mesura suprema o ínfima de la sèrie temporal: $\text{seg}_S(\sup S) = (+\infty,\infty)$ i $\text{ant}_S(\inf S) = (-\infty,\infty)$.

Concatenació

La concatenació és una operació que uneix dues seqüències amb els elements de la primera seqüència seguits pels de la segona. Així doncs, la concatenació de les seqüències té un sentit semblant al que la unió té en els conjunts.

Per a les sèries temporals, per tal que l'operació de concatenació uneixi amb ordre els operands, cal tenir en compte l'interval que ocupa cada sèrie temporal segons el seu atribut de temps. És a dir, la concatenació de dues sèries temporals consisteix a unir la part de la segona sèrie temporal que no està inclosa en el rang temporal de la primera.

Per a poder concatenar dues sèries temporals cal que ambdues tinguin la mateixa estructura, de la mateixa manera que ja s'ha vist amb l'operació d'unió.

Definició 4.38 (Concatenació)**.** *Siguin S_1 i S_2 dues sèries temporals en què* $\operatorname{dom} S_1 = \operatorname{dom} S_2$*, la concatenació de les dues sèries temporals, $S_1||S_2$, és una sèrie temporal que conté totes les mesures de S_1 i les mesures de S_2 que no intersequen en l'interval de S_1: $S_1||S_2 = S_1 \cup (S_2 - S_2[T(\inf S_1), T(\sup S_1)])$.*

Propietats

- La concatenació no és commutativa

4.2.3. Funció temporal

Atenent al fet que una sèrie temporal pot representar-se com una funció temporal cal definir operacions per a tractar convenientment aquesta naturalesa. En aquest apartat definim aquestes operacions com una redefinició de les bàsiques anteriors, i així poder aplicar-les considerant una sèrie temporal com una funció temporal.

A l'apartat 4.3.2 es detalla més el concepte de representació en funció temporal d'una sèrie temporal i s'ofereixen exemples de diversos mètodes de representació. Les operacions definides a continuació han de ser contextualitzades per a un mètode de representació particular. A causa d'això indiquem cada operació de funció temporal amb un superíndex, per exemple r, que indica que el nom r del mètode de representació usat.

Interval temporal

Sigui S una sèrie temporal i $[s,t]$ un interval de temps, per una banda s'ha definit l'interval sobre la seqüència d'una sèrie temporal $S(s,t)$ (v. definició 4.36) i per altra banda la sèrie temporal pot tenir un mètode de representació r que permet calcular la funció temporal de la sèrie temporal $S(x)^r$ (v. definició 4.44) on $x \in \mathcal{T}$. Per seleccionar un interval temporal cal tenir en compte tant l'interval sobre la seqüència com la funció temporal de la sèrie temporal que pot incloure noves mesures al resultat.

Definició 4.39 (Interval temporal)**.** *Sigui S una sèrie temporal, $[s,t]$ un interval de temps i r un mètode de representació, l'interval temporal $S[s,t]^r$ és una sèrie temporal amb les mesures que són dins del rang temporal de l'interval $[s,t]$ segons marca la funció de representació: $S[s,t]^r \equiv S(x)^r$ per tot $x \in [s,t]$*

Aquesta és una definició genèrica difícil d'implementar, per tant per a cada mètode de representació cal interpretar una operació d'interval temporal. Més endavant, un cop haguem profunditzat el concepte de mètode de representació, oferirem exemples d'intervals temporals particularitzats per mètodes de representació (v. secció 4.3.2)

Propietats de l'interval temporal:

- Sigui t un instant de temps, l'interval temporal $S[t,t]^r$ és equivalent a la funció temporal de la sèrie temporal avaluada en aquest instant: $S[t,t]^r = \{(t, S(t)^r)\}$.

Selecció temporal

La selecció temporal d'una sèrie temporal permet seleccionar, en el context d'una representació, un conjunt d'instants de temps determinats. Així, també es pot utilitzar aquesta operació per a canviar la resolució d'una sèrie temporal.

Definició 4.40 (Selecció temporal)**.** *Sigui S una sèrie temporal, $I = \{t_0, t_1, \ldots, t_n\}$ un conjunt d'instants de temps i r un mètode de representació. La selecció temporal, notada com $S[I]^r$, és una sèrie temporal que conté mesures amb els temps d'I segons marca el mètode de representació: $S[I]^r = S[t_0,t_0]^r \cup S[t_1,t_1]^r \cup \cdots \cup S[t_n,t_n]^r$.*

Propietats de la selecció temporal:

- El cardinal de la sèrie temporal resultant és el mateix que el del conjunt d'instants de temps $|S[I]^r| = |I|$

Concatenació temporal

La concatenació temporal és l'operació de concatenació que té en compte la representació de les sèries temporals. És a dir, la concatenació temporal de dues sèries temporals uneix la part de la segona sèrie temporal que no està inclosa en l'interval temporal de la primera.

Definició 4.41 (Concatenació temporal)**.** *Siguin S_1 i $S_2\}$ dues sèries temporals i r un mètode de representació, la concatenació temporal de les dues sèries temporals, $S_1||^r S_2$, és una sèrie temporal que conté les mesures de S_1 i les mesures de S_2 que no intersequen en l'interval temporal de S_1: $S_1||^r S_2 = S_1[s,t]^r \cup S_2[-\infty, s]^r \cup S_2[t, +\infty]^r$ on $s = T(\inf S_1)$ i $t = T(\sup S_1)$.*

Propietats de la concatenació temporal:

- No commutativa

Junció temporal

La junció temporal de dues sèries temporals és la junció que té en compte la representació de les sèries temporals. És a dir, la junció temporal de dues sèries temporals ajunta parelles de mesures seleccionant el mateix atribut de temps en ambdues sèries temporals.

Definició 4.42 (Junció temporal)**.** *Siguin S_1 i S_2 dues sèries temporals i r un mètode de representació, la junció temporal de les dues sèries temporals, $S_1 \bowtie^r S_2$, és una sèrie temporal multivaluada que ajunta les mesures seleccionant els mateixos temps a cada sèrie temporal segons el mètode de representació: $S_1 \bowtie^r S_2 = \{(x, v, w) | x \in (\Pi_t(S_1) \cup \Pi_t(S_2)) \wedge (x, v) \in S_1[x,x]^r \wedge (x, w) \in S_2[x,x]^r\}$*

Propietats de la junció temporal:

- El cardinal resultant és $|S_1 \bowtie^r S_2| \leq |S_1| + |S_2|$
- És commutativa; tenint en compte que els atributs tenen nom i per tant l'ordre no importa.

També es defineix l'operació de semijunció temporal que és una junció no commutativa on la primera sèrie temporal marca el vector de temps de junció.

Definició 4.43 (Semijunció temporal)**.** *Sigui S_1 i S_2 dues sèries temporals i r un mètode de representació, la semijunció temporal de les dues sèries temporals, $S_1 \ltimes^r S_2$, és una sèrie temporal multivaluada que ajunta les mesures de la primera sèrie temporal a les mesures de la segona segons el mètode de representació: $S_1 \ltimes^r S_2 = S_1 \bowtie^r S_2[\Pi_t(S_1)]^r$.*

Propietats de la semijunció temporal:

- El cardinal resultant és $|S_1 \ltimes^r S_2| = |S_1|$.
- No és commutativa.

4.3. Propietats de les sèries temporals

En el model de SGST descrit anteriorment, les sèries temporals tenen una única estructura i uns operadors genèrics definits per un model lògic independentment del context on s'apliquin. No obstant això, en la seva aplicació, les sèries temporals tenen associat un context, és a dir que els valors prenen un determinat significat. En aquesta secció avaluem les propietats que poden prendre les sèries temporals quan es troben en uns determinats contexts. En concret s'avaluen:

- Trets semàntics: el significat que tenen en el seu àmbit d'aplicació
- Grafs i representacions: les visualitzacions i interpretacions possibles
- Patologies: els casos problemàtics de les sèries temporals

4.3.1. Trets semàntics de les sèries temporals

Una sèrie temporal pot provenir de diferents àmbits i per tant tenir un significat variat. És el que anomenem trets semàntics de les sèries temporals. Entendre el significat que tenen les sèries temporals és important per a determinar quines operacions tenen sentit de ser aplicades i quines no a una sèrie temporal en particular. Així Segev i Shoshani [110] anomenen comportament semàntic (*semantic behavior*) a la varietat de formes que pot prendre una sèrie temporal segons l'àmbit on s'apliquen.

A continuació estudiem els tres semàntics de les sèries temporals segons l'origen; és a dir segons com s'han adquirit amb l'aparell de mesura, quina continuïtat té el temps d'adquisició o de si provenen d'un model d'intervals de temps.

Aparell de mesura

Un primer tret semàntic de les sèries temporals és deu al mètode d'adquisició segons l'aparell de mesura. Segons Proakis i Manolakis [104, cap. 1] els senyal discrets tenen principalment dos orígens: l'adquisició mitjançant el mostreig d'un senyal continu cada cert període de temps o l'adquisició que prové d'una acumulació o d'un comptatge d'esdeveniments en un cert període de temps. Aquest tret semàntic descrit fa referència a l'eix del temps. Pel que fa a l'eix de valors es pot distingir entre una adquisició contínua o discreta, procés que s'anomena quantització del senyal, tot i que sovint per simplificar no es té en compte [104]. En el nostre cas, per tal de no restringir el model a cap domini de valors en concret (v. apartat 4.1.2), no avaluem el significat de la quantització en les sèries temporals perquè només afecta a les operacions que treballen amb l'atribut de valor.

S_m

t	v
0	0
2	1
3	1,5
4	0
5	2
8	2

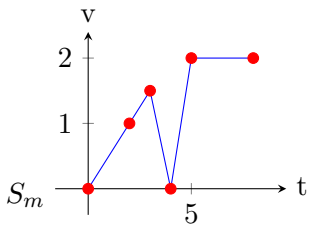


Figura 4.10.: Magnitud física

Així doncs, classifiquem els trets semàntics d'una sèrie temporal en magnitud física o en comptador de magnitud física segons si l'aparell de mesura és continu o d'acumulació. Aquest tret semàntic cal tenir-lo en compte ja que quan s'adquireix una magnitud es mostra el seu estat instantani actual, mentre que quan s'adquireix un comptador ofereix informació de la seva variació respecte a l'adquisició anterior.

Les sèries temporals amb tret semàntic de magnitud ofereixen la informació en una sola forma: l'estat de la variable. A la figura 4.10 es mostra una sèrie temporal d'exemple: $S_m = \{(0,0),(2,1),(3,1{,}5),(4,0),(5,2),(8,2)\}$ on els valors mostren l'estat de la magnitud física en els instants de temps mesurats. De manera general, el valor d'una magnitud física és defineix com el producte d'un valor numèric per una unitat.

Les sèries temporals amb tret semàntic de comptador poden oferir la informació en diverses formes : en valor absolut, en valor relatiu i en velocitat. A la figura 4.11 es mostren aquestes formes d'informació mitjançant tres sèries temporals d'exemple provinents del mateix comptatge:

- $S_a = \{(0,0),(2,1),(4,3),(5,4),(8,10)\}$ amb els valors absoluts del comptador
- $S_\Delta = \{(0,0),(2,1),(4,2),(5,1),(8,6)\}$ amb els valors relatius del comptador, on $S_\Delta = \{(0,0)\} \cup \text{increments}(S_a)$
- $S_v = \{(0,0),(2,0{,}5),(4,1),(5,1),(8,2)\}$ amb els valors de velocitat del comptador, on $S_v = S_\Delta / \text{increments}(\text{duplica\_t}(S_\Delta))$ (v. exemple 4.13 i exemple 4.17) i les unitats dels valors són de *unitat/unitat de temps*.

A la figura 4.11b es mostra la sèrie temporal S_a, en aquest cas es tracta d'un comptador monòton creixent. S'hi mostra la interpolació lineal del valor acumulat que va prenent la variable mesurada. A partir del valor absolut es calculen els increments del comptador S_Δ, és a dir la quantitat relativa de comptatge per cada període. A la figura 4.11c es mostra la sèrie temporal S_Δ indicant en línia discontínua la tendència dels increments, és a dir la interpolació lineal del comptatge absolut de manera similar a com s'ha fet per la mesura de la magnitud de la figura 4.10. A la figura 4.11e també es mostra S_Δ però indicant la interpolació lineal que va prenent

S_a

t	v
0	0
2	1
4	3
5	4
8	10

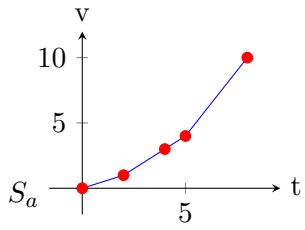
S_Δ

t	v
0	0
2	1
4	2
5	1
8	6

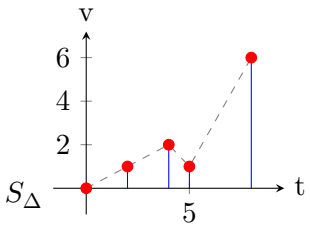
S_v

t	v
0	0
2	0,5
4	1
5	1
8	2

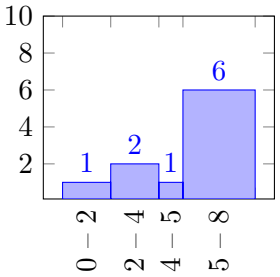
(a) Taules de valors



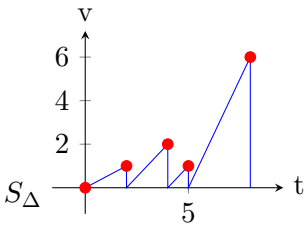
(b) Valor absolut



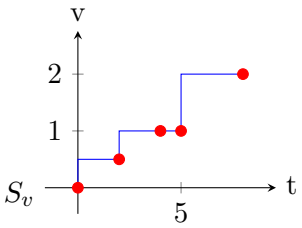
(c) Increments



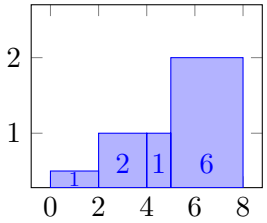
(d) Gràfic de barres



(e) Valor relatiu



(f) Velocitat



(g) Histograma

Figura 4.11.: Formes d'un comptador monòton

la variable mesurada relativament, és a dir amb el pas per zero com si a cada lectura s'inicialitzés de nou el valor acumulat del comptador. A la figura 4.11f es mostra la sèrie temporal S_v, que es correspon amb els pendents lineals del gràfic de valor relatiu, és a dir la velocitat mitjana de comptatge a cada interval de temps.

El significat del gràfic de valor absolut i el d'increments es visualitza bé en el gràfic de barres de la figura 4.11d en què l'eix vertical mostra freqüència i l'eix horitzontal mostra la quantitat que ha augmentat el comptador a cada interval. En canvi el gràfic de valor relatiu i el de velocitat es visualitzen més bé en el gràfic d'histograma de la figura 4.11g en què l'eix vertical mostra densitat de freqüència, l'eix horitzontal mostra la velocitat del comptatge i l'àrea de les barres equival a la quantitat comptada.

En conclusió, un comptador pot oferir les dades en qualsevol de les tres formes i a partir d'una d'aquestes es poden calcular les altres. També es possible d'aplicar les mateixes operacions a les magnituds, és a dir calcular-ne els increments o la velocitat mitjana a cada cert interval de temps, tenint en compte que aleshores es perd la referència absoluta llevat que s'emmagatzemi a part. La diferència principal d'ambdós rau en el fet que l'objectiu de les magnituds és observar el valor instantani d'una variable, i per tant se sol obtenir en aquesta forma, mentre que l'objectiu dels comptadors és precisament el de mesurar els comptatges totals de les variables en els intervals de temps seleccionats, i això es pot mostrar en les tres formes descrites. És a dir, el trets semàntics de la sèrie temporal depenen de l'adquisició per part de l'aparell de mesura: una magnitud ben mostrejada té informació suficient per a calcular-ne els comptatges mentre que un comptador inherentment mesura els totals però perd la noció del valor instantani.

Altrament, tot i que els comptadors solen presentar un comportament monòton, creixent o decreixent, també poden ser dobles i tenir increments tant positius com negatius. Un efecte que també cal tenir en compte dels comptadors, sobretot dels monòtons, és el fons d'escala, que és aquell valor que quan l'assoleixen es torna a comptar des de l'inici i per tant, en aquest interval de mesura, tenen un comportament semblant al de valor relatiu.

Exemple 4.19 (Sèrie temporal amb trets de magnitud i de comptador)**.** Un exemple de sèrie temporal amb tret semàntic de magnitud és la potència instantània elèctrica i un exemple de comptador és la quantitat d'energia elèctrica. Així doncs mesurant electricitat poden aparèixer sèries temporals amb els trets semàntics següents:

- Magnitud: potència instantània elèctrica. Per exemple la sèrie S_m amb unitats als valors de watts.
- Comptador de valor absolut: energia elèctrica. Per exemple la sèrie S_a amb unitats als valors de watts×hora o bé de joules.

- Comptador de valor relatiu: energia elèctrica amb posada a zero a cada lectura. Per exemple la sèrie S_Δ amb unitats als valors de watts×hora o bé de joules.
- Comptador de velocitat: potència mitjana elèctrica. Per exemple la sèrie S_v amb unitats als valors de watts o bé de watts×hora/hora.

Si la potència instantània és periòdica es pot mostrejar correctament per a saber el total d'energia elèctrica, si no és periòdica difícilment es pot determinar la quantitat amb exactitud. En canvi, mentre que un aparell de comptatge mesura exactament la quantitat d'energia, no es pot recuperar la potència instantània si no es coneix quina forma té.

Continuïtat del temps

Un segon tret semàntic de les sèries temporals també es deu al mètode d'adquisició però en aquest cas segons la continuïtat del temps.

Furia et al. [51] distingeixen entre temps dens (*dense time*) i temps discret (*discrete time*) segons si entre dos instants de temps n'hi pot haver un de tercer o si els instants de temps son punts aïllats. Un exemple de domini del temps com a conjunt dens són els reals, $\mathcal{T} = \bar{\mathbb{R}}$, i com a conjunt discret són els naturals, $\mathcal{T} = \mathbb{N}$. Kopetz [74, cap. 3] descriu conceptes similars tot i que els anomena *dense time* i *sparse time* i ho fa des del punt de vista d'esdeveniments de temps real. Així fa notar que els esdeveniments es produeixen amb *sparse time* quan hi ha un sistema que controla el rellotge, p.ex. un processador, però que els esdeveniments que són externs a aquest control sempre es produeixen amb *dense time*.

Així, segons la perspectiva del temps com a sistema de coordenades de la definició 4.1, el temps dens infereix als instants de temps un significat de punt temporal (*timestamp*) mentre que el temps discret infereix un significat de lapse de temps (*time span*) a vegades també anomenat *chronon*.

Un cas particular de sèries temporals amb trets semàntics de temps discret són les que tenen trets de seqüència. Les sèries temporals com a seqüències són habituals en l'anàlisi de sèries temporals ja que són una manera simple d'observar un conjunt de dades ordenades sense determinar-ne exactament la distància temporal entre elles ni la posició absoluta en el sistema de coordenades.

Exemple 4.20 (Sèrie temporal amb tret de seqüència densa)**.** Un cas que observem amb tret semàntic de seqüència són les sèries temporals que recullen dades per intervals de temps. Per exemple, una sèrie temporal que recull les temperatures mitjanes de cada mes $S = \{(\text{gen99}, 10), (\text{feb99}, 11), (\text{març99}, 15), \dots\}$. En aquest cas el temps és discret amb el domini $\mathcal{T} = \{\text{gen99}, \text{feb99}, \text{març99}, \dots\}$ i cada instant de temps es pot observar com un lapse que dura tot el mes. Això no obstant, aquestes sèries temporals s'expressen amb més exactitud amb temps dens, per exemple

en el cas de les mitjanes s'hauran adquirit a final de mes i per tant els instants de temps poden ser més precisos $S = \{(\text{31-gen-99 } 23.59, 10), (\text{28-feb-99 } 23.59, 11), (\text{31-març-99 } 23.59, 15), \dots\}$.

Exemple 4.21 (Sèrie temporal amb tret de seqüència discreta)**.** Un altre cas que observem amb trets de seqüència són les sèries temporals que recullen agregacions d'intervals de temps no continus. Per exemple una sèrie temporal que recull les temperatures mitjanes dels mesos d'octubre $S = \{(\text{oct98}, 15), (\text{oct99}, 10), (\text{oct00}, 13), \dots\}$. En aquest cas el temps és discret amb el domini $\mathcal{T} = \{\text{oct98}, \text{oct99}, \text{oct00}, \dots\}$ i cada instant de temps es pot observar com un lapse que dura tot un mes d'octubre però no hi ha continuïtat del temps entre els mesos. Així doncs, aquestes sèries temporals presenten més dificultats a l'hora d'expressar-se amb temps dens, per exemple com en el cas anterior es pot precisar amb els instants de temps $S = \{(\text{31-oct-98}, 15), (\text{31-oct-99},10), (\text{31-oct-00},13), \dots\}$ o bé refermar-ho amb valors indefinits $S = \{(\text{30-set-98}, \infty), (\text{31-oct-98}, 15), (\text{1-nov-98}, \infty), (\text{30-set-99}, \infty), (\text{31-oct-99},10), \dots\}$, però cal anar en compte amb les operacions que s'hi apliquin ja que poden no adequar-se a la semàntica si no tenen en compte aquesta discontinuïtat del temps.

Exemple 4.22 (Sèrie temporal amb tret de seqüència agregada)**.** Un cas semblant als dos exemples anteriors amb trets de seqüència són les sèries temporals que mostren agregacions de dades per intervals de temps. Per exemple, una sèrie temporal que recull la mitjana de les temperatures de cada mes $S = \{(\text{gen}, 10), (\text{feb}, 11), \dots, (\text{des}, 15)\}$. En aquest cas el temps és discret amb el domini $\mathcal{T} = \{\text{gen}, \text{feb}, \dots, \text{des}\}$ i el lapse dura tot un mes però sense cap posició absoluta en el sistema de coordenades de temps. Aquest cas no es pot expressar amb temps dens.

Encara que aquest i casos semblants compleixen les propietats de sèries temporals, la interpretació del temps és dificultosa i pot resultar més còmode tractar-los com a casos genèrics dels SGBDR on la interpretació de l'atribut de temps com a característica agrupada serà més apropiada.

Intervals temporals

Finalment, hi ha dades que s'anoten en intervals de temps en comptes d'instants de temps. Aquests casos pertanyen al model d'intervals de temps, el qual és diferent del de sèries temporals. Algunes dades en intervals temporals es poden expressar perfectament en sèries temporals però altres no resulta tan còmode. Vegem-ne alguns exemples.

Sense definir amb detall el model d'intervals temporals [33], assumim que en el model d'intervals l'estat d'una variable s'indica amb el conjunt $I = \{([t_0, t_1), v_0), ([t_1, t_2), v_1), \dots\}$ que expressa que en l'interval de temps $[t_0, t_1)$ la variable mesurada té el valor v_0, en l'interval de temps següent $[t_1, t_2)$ té el valor v_1, etc.

Exemple 4.23 (Intervals temporals del tipus d'interès)**.** Un primer exemple és el tipus d'interès que fixa el Banc Central Europeu. En aquest cas es fixen valors cadascun dels quals és vàlid durant un determinat interval de temps. Un exemple per a un tipus d'interès expressat en intervals temporals és $I = \{([\text{2012-07-11}, \text{2013-05-08}), \text{0,75}), ([\text{2013-05-08}, \text{2013-11-13}), \text{0,50}), ([\text{2013-11-13}, \text{2014-06-11}), \text{0,25}), ([\text{2014-06-11}, +\infty), \text{0,15})\}$ on els valors tenen unitats de tant per cent, 2012-07-11 és el primer instant conegut, i 2014-06-11 és el darrer instant fixat i per tant el valor 0,15 és d'aplicació fins a un nou canvi. Es pot expressar fàcilment en sèrie temporal $S = \{(\text{2012-07-11}, \text{0,75}), (\text{2013-05-08}, \text{0,50}), (\text{2013-11-13}, \text{0,25}), (\text{2014-06-11}, \text{0,15})\}$, i acompanyar-ho amb una representació adequada com la ZOH (v. definició 4.48).

Exemple 4.24 (Intervals temporals d'una agenda)**.** Un segon exemple és una agenda, la qual fixa tasques en intervals de temps futurs. Per exemple expressat en intervals temporals $I = \{([7, 8), \text{transport}), ([8, 15), \text{feina}), ([17, 18), \text{dentista})\}$. També com en el cas anterior es pot expressar en sèrie temporal $S = \{(7, \text{transport}), (8, \text{feina}), (15, \infty), (17, \text{dentista}), (18, \infty)\}$. Tot i així, per a aquest exemple la sèrie temporal no és còmode i és millor el model d'intervals temporals, el qual està més orientat a expressar el temps de validesa de cada tasca.

4.3.2. Graf i funció temporal de representació

Una sèrie temporal pot representar discretament una funció temporal (original), és a dir una funció que depèn de la variable temps. Una sèrie temporal pot representar-se com una funció temporal, és a dir obtenir nous valors a partir de la sèrie temporal segons una funció. Aquesta nova col·lecció de valors s'anomena graf d'una funció (*graph of a function*), en el sentit de gràfic (*plot*) que cal no confondre amb els grafs d'arestes i vèrtexs (*vertex-edge graph*). Així, una sèrie temporal pot representar-se com una funció que descrigui o s'aproximi a com era la funció temporal original representada. En aquesta secció, però, descrivim com poden representar-se independentment que s'ajustin a la funció temporal original.

Per a calcular el graf d'una sèrie temporal es necessita una funció de representació. Mentre que la funció que permet canviar d'una funció contínua a una sèrie temporal s'anomena procés d'adquisició o mostreig, la funció que permet canviar d'una sèrie temporal a una funció contínua l'anomenem funció de representació. Així doncs, donada una sèrie temporal es poden definir funcions amb el temps com a variable que calculin nous valors a partir de les mesures emmagatzemades.

Definició 4.44 (Funció de representació)**.** *Sigui S una sèrie temporal, es defineix $S(t)$ com la funció de representació de la sèrie temporal contínua al llarg del temps, $t \in \mathcal{T}$. És a dir que per cada instant de temps la funció pren un valor i per tant és una funció $\mathcal{T} \to \mathcal{V}$.*

Atenent a les operacions de càlcul que es facin per a obtenir $S(t)$ diem que hi ha diverses funcions de representació. Així per a cada funció de representació indicarem a quina ens referim amb un superíndex $S^r(t)$ on r és el nom d'una funció de representació de la sèrie temporal.

Definició 4.45 (Graf d'una sèrie temporal)**.** *Sigui $S(t)$ la funció de representació d'una sèrie temporal i $\mathcal{T}$ un domini del temps, es defineix el graf de la sèrie temporal,* graf $S(t)$*, com un conjunt de parells ordenats $(t, S(t))$ de manera que* graf $S(t) = \{(t, S(t))|t \in \mathcal{T}\}$.

La utilitat de les funcions de representació és diversa i per això les operacions de càlcul poden ser qualssevol. Una funció de representació es pot utilitzar per a interpolar valors d'una sèrie temporal però també per a extrapolar-los i fer prediccions o per a canviar la resolució de la sèrie temporal. També es pot utilitzar com a tècnica d'aproximar la sèrie temporal a la funció original; és a dir trobar una funció de representació que es correspongui amb la funció contínua que més s'aproxima a la funció temporal original.

El lligam entre una sèrie temporal i la seva representació no és fix; és a dir que donada una sèrie temporal es pot representar amb una funció o amb una altra segons convingui. Encara que en alguns àmbits, per exemple a teoria del senyal o en algunes aplicacions de l'anàlisi de sèries temporals, l'objectiu és cercar la parella de sèrie temporal i representació que més s'aproxima a la funció contínua original; en altres àmbits, com per exemple el del model multiresolució que definim posteriorment, l'objectiu està més orientat a utilitzar representacions de les sèries temporals segons els càlculs que es volen fer o segons la naturalesa que s'assumeixi de la sèrie temporal.

No obstant això, les funcions de representació s'han d'utilitzar amb criteri. Per exemple, els trets semàntics de la sèrie temporal determinen el significat de les operacions que es calculen per a fer la representació, per tant l'aplicació de qualsevol funció de representació a una sèrie temporal pot donar resultats incoherents. També caldrà tenir en compte la semàntica si s'apliquen càlculs successius a una sèrie temporal seguint diferents funcions de representació. Això no obstant, en la formalització del model de funcions de representacions no definim cap criteri en concret per a, així, donar llibertat en les possibilitats de càlcul amb les sèries temporals. En la secció 2.1 s'ha comentat breument la diversitat en els algoritmes de representació que s'utilitzen per a l'anàlisi de sèries temporals.

A continuació definim diverses funcions de representació per a exemplificar-ne l'ús. Les agrupem per algunes de les seves característiques i que, d'alguna manera, formen famílies de funcions de representació; així de cada una en definim les més representatives. Com a mostra de la diversitat, exemplifiquem tres grans famílies de funcions: parcials, a trossos i aproximacions. En les definicions de les funcions a trossos oferim, quan es pugui, dues expressions equivalents: una com a sumatori

S

t	v
3	1
4	3
6	2
9	1

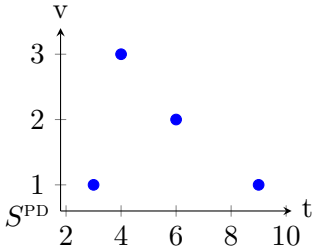


Figura 4.12.: Taula d'una sèrie temporal S i graf $S^{\text{PD}}(t)$

de subfuncions contínues, que ajuda a comprendre'n el significat, i una altra com a taula de pertinences, que utilitza l'àlgebra del model de SGST. Finalment, un cop presentats els diversos exemples de funcions de representació, explicitem l'ús que tenen en les operacions dels SGST.

Funcions parcials

Les funcions parcials no són totalment contínues en el temps sinó que són funcions no definides en alguns valors de temps; contràriament les funcions definides per tots els valors de temps s'anomenen funcions totals. Així, en aquests casos parcials, no hi ha una interpretació clara de què hi ha entre dues mesures. A continuació ho exemplifiquem amb una representació que anomenem parcial discreta (PD).

Definició 4.46 (Funció de representació parcial discreta)**.** *Sigui S una sèrie temporal, es defineix $S^{\text{PD}}(t)$ com la funció de representació parcial discreta de la sèrie temporal, $\forall m \in S$:*

$$S^{\text{PD}}(t) = \begin{cases} V(m) & \textit{si } t = T(m) \\ \textit{no definit} & \textit{altrament} \end{cases}$$

Aquest és un cas especial de funció de representació perquè permet que el graf de la sèrie temporal sigui equivalent a les mesures de la sèrie temporal. Sigui $S = \{m_0, \ldots, m_k\}$ una sèrie temporal aleshores graf $S^{\text{PD}}(t) \equiv \{m_0, \ldots, m_k\}$.

Exemple 4.25 (Sèrie temporal amb representació parcial discreta)**.** Sigui la sèrie temporal $S = \{(3,1),(4,3),(6,2),(9,1)\}$, el graf de la representació parcial discreta és graf $S^{\text{PD}}(t) = \{(3,1),(4,3),(6,2),(9,1)\}$, el qual es mostra a la figura 4.12.

A impulsos

Una família de funcions contínues que recorda a la funció discreta són les funcions d'impulsos (*impulse train function*), tot i que pertanyen a les funcions definides contínues a trossos. A continuació ho exemplifiquem amb una representació que

S

t	v
3	1
4	3
6	2
9	1

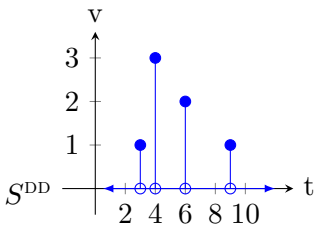


Figura 4.13.: Taula d'una sèrie temporal S i graf $S^{\text{DD}}(t)$

anomenem delta de Dirac (DD) perquè es basa en la funció homònima, la qual val zero a tot arreu excepte en el punt zero.

Definició 4.47 (Funció de representació delta de Dirac)**.** *Sigui S una sèrie temporal i $\mathcal{T}$ el domini del temps, es defineix $S^{\text{DD}}(t)$ com la funció de representació delta de Dirac al llarg del temps, $\forall m \in S$:*

$$
\begin{aligned}
S^{\text{DD}}(t) &= \\
&= \sum_{t \in \mathcal{T}} V(m)\delta(t - T(m)) : \delta(t) = \begin{cases} 1 & si\ t = 0 \\ 0 & altrament \end{cases} \\
&= \begin{cases} V(m) & si\ t = T(m) \\ 0 & altrament \end{cases}
\end{aligned}
$$

.

Exemple 4.26 (Sèrie temporal amb representació delta de Dirac)**.** Sigui la sèrie temporal $S = \{(3,1),(4,3),(6,2),(9,1)\}$, la seva representació DD és $S^{\text{DD}}(t) = 0 + 1\delta(t-3)+3\delta(t-4)+2\delta(t-6)+1\delta(t-9)$. El graf d'aquesta representació, graf $S^{\text{DD}}(t)$, es mostra a la figura 4.13.

A trossos constants

Una altra família de funcions contínues són les que es basen en funcions definides constants a trossos *piecewise constant functions*). A continuació ho exemplifiquem amb quatre representacions basades en la funció graó (*step function* o *staircase function*) atenent a quatre de les possibles continuïtats en els intervals de temps.

A les definicions següents s'utilitza la notació de funció característica $I_A(t)$ per a indicar quan un instant de temps pertany a un determinat interval de temps:

$$
I_A(t) = \begin{cases} 1 & \text{si } t \in A \\ 0 & \text{altrament} \end{cases}
$$

En primer lloc, definim una representació en base a funcions graó contínues per la dreta. L'anomenem representació *zero-order hold* (ZOH) a causa de la semblança

que té amb el model utilitzat en teoria del senyal per a reconstruir senyals, el qual consisteix en mantenir constant cada valor fins al proper.

Definició 4.48 (Funció de representació *zero-order hold*)**.** *Sigui S una sèrie temporal i $\mathcal{T}$ el domini del temps, es defineix $S^{\text{ZOH}}(t)$ com la funció de representació* zero-order hold *al llarg del temps, $\forall m \in S$:*

$$\begin{aligned} S^{\text{ZOH}}(t) = & \\ &= \sum_{t \in \mathcal{T}} V(m) I_{\left[T(m), T(\text{seg}(m))\right)}(t) \\ &= \begin{cases} 0 & si\ t < T(\min(S)) \\ V(m) & si\ t \in \left[T(m), T(\text{seg}(m))\right) \end{cases} \end{aligned}$$

.

En segon lloc, definim una representació en base a funcions graó contínues per l'esquerra. L'anomenem representació *zero-order hold* cap enrere (ZOHE) perquè consisteix en mantenir constant cada valor fins al predecessor. Una representació similar s'utilitza a RRDtool [96].

Definició 4.49 (Funció de representació *zero-order hold* cap enrere)**.** *Sigui S una sèrie temporal i $\mathcal{T}$ el domini del temps, es defineix $S^{\text{ZOHE}}(t)$ com la funció de representació* zero-order hold *cap enrere al llarg del temps, $\forall m \in S$:*

$$\begin{aligned} S^{\text{ZOHE}}(t) = & \\ &= \sum_{t \in \mathcal{T}} V(m) I_{\left(T(\text{ant}(m)), T(m)\right]}(t) \\ &= \begin{cases} 0 & si\ t > T(\max(S)) \\ V(m) & si\ t \in \left(T(\text{ant}(m)), T(m)\right] \end{cases} \end{aligned}$$

.

En tercer lloc, definim una representació en base a funcions graó contínues per la dreta centrades en l'interval. L'anomenem representació *zero-order hold* centrada en l'interval (ZOHC).

Definició 4.50 (Funció de representació *zero-order hold* centrada en l'interval)**.** *Sigui S una sèrie temporal i $\mathcal{T}$ el domini del temps, es defineix $S^{\text{ZOHC}}(t)$ com la funció de representació* zero-order hold *centrada en l'interval al llarg del temps, $\forall m \in S$:*

$$\begin{aligned} S^{\text{ZOHC}}(t) = & \\ &= \sum_{t \in \mathcal{T}} V(m) I_{\left[\frac{T(\text{ant}(m))+T(m)}{2}, \frac{T(m)+T(\text{seg}(m))}{2}\right)}(t) \\ &= V(m) : t \in \left[\frac{T(\text{ant}(m)) + T(m)}{2}, \frac{T(m) + T(\text{seg}(m))}{2}\right) \end{aligned}$$

.

S

t	v
3	1
4	3
6	2
9	1

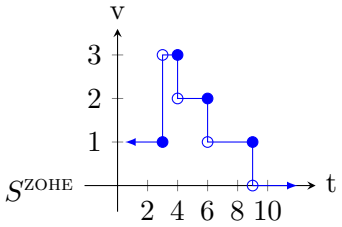


Figura 4.14.: Taula d'una sèrie temporal S i graf $S^{\text{ZOHE}}(t)$

En quart lloc, representem la sèrie temporal en base a la funció rectangular. La funció rectangular és un cas especial de les funcions graó en la qual s'especifiquen valors simètrics per als punts de discontinuïtat. Definim el rectangle amb manteniment del valor cap a la dreta, de manera similar al ZOH, tot i que també és possible definir altres translacions del rectangle com s'ha fet per la funció graó en els casos del ZOHE i del ZOHC.

Definició 4.51 (Funció de representació rectangular)**.** *Sigui S una sèrie temporal i $\mathcal{T}$ el domini del temps, es defineix $S^{rect}(t)$ com la funció de representació rectangular al llarg del temps, $\forall m \in S$:*

$$S^{rect}(t) =$$

$$= \sum_{t \in \mathcal{T}} V(m)\,\text{rect}(t) : \text{rect}(t) = \begin{cases} 1 & si\ t \in (T(m), T(\text{seg}(m))) \\ \frac{1}{2} & si\ t = T(m) \vee t = T(\text{seg}(m)) \\ 0 & altrament \end{cases}$$

$$= \begin{cases} 0 & si\ t < T(\min(S)) \\ V(m) & si\ t \in (T(m), T(\text{seg}(m))) \\ \frac{V(m)+V(\text{ant}(m))}{2} & si\ t = T(m) \wedge t > T(\min(S)) \\ \frac{V(m)}{2} & si\ t = T(\min(S)) \end{cases}$$

.

Exemple 4.27 (Sèrie temporal amb representació ZOHE)**.** Sigui la sèrie temporal $S = \{(3,1),(4,3),(6,2),(9,1)\}$, la seva representació ZOHE és $S^{\text{ZOHE}}(t) = 1I_{(-\infty,3]} + 3I_{(3,4]} + 2I_{(4,6]} + 1I_{(6,9]} + 0I_{(9,+\infty)}$. El graf d'aquesta representació, graf $S^{\text{ZOHE}}(t)$, es mostra a la figura 4.14.

Exemple 4.28 (Sèrie temporal amb representació rectangular)**.** Sigui la sèrie temporal $S = \{(3,1),(4,3),(6,2),(9,1)\}$, la seva representació rectangular és $S^{\text{rect}}(t) = 0I_{(-\infty,3)} + 0{,}5I_{[3,3]} + 1I_{(3,4)} + 2I_{[4,4]} + 3I_{(4,6)} + 2{,}5I_{[6,6]} + 2I_{(6,9)} + 1{,}5I_{[9,9]} + 1I_{(9,+\infty)}$. El graf d'aquesta representació, graf $S^{\text{rect}}(t)$, es mostra a la figura 4.15.

S

t	v
3	1
4	3
6	2
9	1

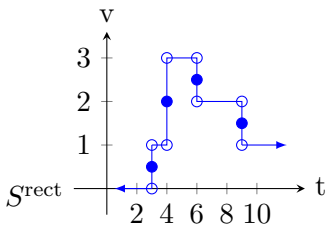


Figura 4.15.: Taula d'una sèrie temporal S i graf $S^{\text{rect}}(t)$

A trossos lineals

Una família de funcions d'un ordre superior a les de trossos constants són les que es basen en funcions definides lineals a trossos (*piecewise linear functions*). A continuació ho exemplifiquem amb una representació basada en la funció triangular (*triangular function*). Seguint l'analogia amb la teoria del senyal, l'anomenem representació *first-order hold* (FOH), el qual consisteix en interpolar linealment cada valor fins al proper.

La definició del FOH amb funcions matemàtiques contínues es construeix a partir de la funció triangular $\text{tri}(t)$:

$$\text{tri}(t) = \begin{cases} 1 - |t| & \text{si } |t| < 1 \\ 0 & \text{altrament} \end{cases}$$

Per al cas particular de sèries temporals regulars (v. definició 4.59), es pot utilitzar directament la funció triangular general. Sigui R una sèrie temporal regular amb període p i sigui $\mathcal{T}$ el domini del temps, es defineix $R^{\text{FOH}}(t)$ com la funció de representació *first-order hold* al llarg del temps, $\forall m \in R$:

$$R^{\text{FOH}}(t) = \sum_{t \in \mathcal{T}} V(m)\, \text{tri}\left(\frac{t - T(m)}{p}\right)$$

Per al cas general, tant per a sèries temporals regulars com per a no regulars, s'ha de construir a partir de funcions triangulars no simètriques:

$$\text{tri}^{\text{ant}}(t) = \begin{cases} 1 - |t| & \text{si } -1 < t < 0 \\ 0 & \text{altrament} \end{cases} \qquad \text{tri}^{\text{seg}}(t) = \begin{cases} 1 - |t| & \text{si } 0 \leq t < 1 \\ 0 & \text{altrament} \end{cases}$$

Definició 4.52 (Funció de representació *first-order hold*)**.** *Sigui S una sèrie temporal i $\mathcal{T}$ el domini del temps, es defineix $S^{\text{FOH}}(t)$ com la funció de representació*

S

t	v
3	1
4	3
6	2
9	1

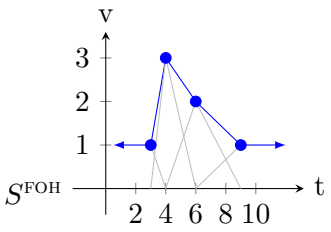


Figura 4.16.: Taula d'una sèrie temporal S i graf $S^{\text{FOH}}(t)$

first-order hold *al llarg del temps,* $\forall m \in S$:

$$
\begin{aligned}
S^{FOH}(t) = & \\
&= \sum_{t \in \mathcal{T}} y_1 \left(\text{tri}^{ant}\left(\frac{t - x_1}{x_1 - x_0}\right) + \text{tri}^{seg}\left(\frac{t - x_1}{x_2 - x_1}\right)\right) \\
&= \begin{cases} V(\min(S)) & si\ t < T(\min(S)) \\ V(\max(S)) & si\ t > T(\max(S)) \\ \frac{y_2 - y_1}{x_2 - x_1}(t - x_1) + y_1 & si\ t \in [x_1, x_2) \wedge t \leq T(\max(S)) \end{cases} \\
&: x_1 = T(m), y_1 = V(m), \\
&\quad m_s = \text{seg}(m), x_2 = T(m_s), y_2 = V(m_s), \\
&\quad m_a = \text{ant}(m), x_0 = T(m_a), y_0 = V(m_a)
\end{aligned}
$$

.

Exemple 4.29 (Sèrie temporal amb representació FOH)**.** Sigui la sèrie temporal $S = \{(3,1),(4,3),(6,2),(9,1)\}$, la seva representació FOH és

$$
\begin{aligned}
S^{\text{FOH}}(t) = & \\
& 1\,\text{tri}^{\text{ant}}\left(\frac{t-3}{3-(-\infty)}\right) + 1\,\text{tri}^{\text{seg}}\left(\frac{t-3}{4-3}\right) + 3\,\text{tri}^{\text{ant}}\left(\frac{t-4}{4-3}\right) + 3\,\text{tri}^{\text{seg}}\left(\frac{t-4}{6-4}\right) \\
& +2\,\text{tri}^{\text{ant}}\left(\frac{t-6}{6-4}\right) + 2\,\text{tri}^{\text{seg}}\left(\frac{t-6}{9-6}\right) + 1\,\text{tri}^{\text{ant}}\left(\frac{t-9}{9-6}\right) + 1\,\text{tri}^{\text{seg}}\left(\frac{t-9}{+\infty-9}\right)
\end{aligned}
$$

El graf d'aquesta representació, graf $S^{\text{FOH}}(t)$, es mostra a la figura 4.16 en què en gris es dibuixa cada funció triangular i en blau la funció resultant del sumatori de totes les triangulars.

D'ajustament de corbes

Una família de funcions diferents als casos anteriors són les funcions d'ajustament de corbes amb polinomis. L'ajustament de corbes (*curve fitting*) és una tècnica que es basa en l'aproximació als punts donats, a diferència de les funcions definides a

trossos anteriors que es basen en interpolacions que passin exactament pels punts donats. A continuació ho exemplifiquem amb una representació basada en l'aproximació lineal per mínims quadrats, és a dir la regressió lineal (*linear regression*), i ho generalitzem per a representacions basades en aproximacions polinomials.

Definim una representació que consisteix en trobar la recta d'ajust per mínims quadràtics a les mesures de la sèrie temporal. L'anomenem representació *lineal* de la sèrie temporal.

Definició 4.53 (Funció de representació lineal)**.** *Sigui $S = \{m_0, \ldots, m_k\}$ una sèrie temporal, es defineix $S^{lineal}(t)$ com la funció de representació de regressió lineal al llarg del temps, $\forall m \in S : S^{lineal}(t) = \alpha + \beta t$ on els valors de la regressió lineal es calculen aplicant:*

$$\begin{bmatrix} \alpha \\ \beta \end{bmatrix} = (A^T A)^{-1} A^T B,$$
$$A = \begin{bmatrix} 1 & T(m_0) \\ \vdots & \vdots \\ 1 & T(m_k) \end{bmatrix}, B = \begin{bmatrix} V(m_0) \\ \vdots \\ V(m_k) \end{bmatrix}$$

Aplicant una equació similar es pot generalitzar el problema a una regressió n-polinomial.

Definició 4.54 (Funció de representació polinomial)**.** *Sigui $S = \{m_0, \ldots, m_k\}$ una sèrie temporal i n el grau del polinomi que es vol ajustar, es defineix $S^{polinomi}(t)$ com la funció de representació de regressió polinomial al llarg del temps, $\forall m \in S : S^{polinomi}(t) = \alpha_0 + \alpha_1 t + \cdots + \alpha_n t^n$ on els valors de la regressió es calculen aplicant:*

$$\begin{bmatrix} \alpha_0 \\ \alpha_1 \\ \vdots \\ \alpha_n \end{bmatrix} = (A^T A)^{-1} A^T B,$$
$$A = \begin{bmatrix} 1 & T(m_0) & \ldots & (T(m_0))^n \\ \vdots & \vdots & \ddots & \vdots \\ 1 & T(m_k) & \ldots & (T(m_k))^n \end{bmatrix}, B = \begin{bmatrix} V(m_0) \\ \vdots \\ V(m_k) \end{bmatrix}$$

Hem exemplificat la família d'ajustament de corbes de forma genèrica per a tota la sèrie temporal, tot i que també és possible aproximar una sèrie temporal a trossos com fan Last et al. [79]. Així, siguin $t_a, t_b, t_c, \ldots$ uns instants de temps que defineixen cadascun dels intervals on es fa l'aproximació, una representació polinomial a trossos $S^{\text{tr-polinomial}}$ tindria la forma

$$S^{\text{tr-polinomial}} = \begin{cases} S[t_a, t_b]^{\text{polinomial}}(t) & \text{si } t_a \leq t < t_b, \\ S[t_b, t_c]^{\text{polinomial}}(t) & \text{si } t_b \leq t < t_c, \\ \ldots \end{cases}$$

S

t	v
3	1
4	3
6	2
9	1

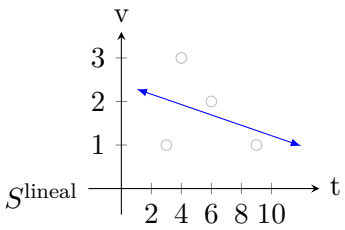


Figura 4.17.: Taula d'una sèrie temporal S i graf $S^{\text{lineal}}(t)$

Exemple 4.30 (Sèrie temporal amb representació lineal)**.** Sigui la sèrie temporal $S = \{(3,1),(4,3),(6,2),(9,1)\}$, la seva representació lineal és $S^{\text{lineal}}(t) = 2{,}405 - 0{,}119t$. El graf d'aquesta representació, graf $S^{\text{lineal}}(t)$, es mostra a la figura 4.17.

Els paràmetres d'aquesta regressió lineal provenen de la resolució del sistema d'equacions següent:

$$S^{\text{lineal}}(t) = \alpha + \beta t$$

$$A = \begin{bmatrix} 1 & 3 \\ 1 & 4 \\ 1 & 6 \\ 1 & 9 \end{bmatrix}, B = \begin{bmatrix} 1 \\ 3 \\ 2 \\ 1 \end{bmatrix}$$

$$\begin{bmatrix} \alpha \\ \beta \end{bmatrix} = (A^T A)^{-1} A^T B = \begin{bmatrix} 2{,}405 \\ 0{,}119 \end{bmatrix}$$

Ús en els operadors de funció temporal

Alguns operadors dels SGST presenten particularitats a causa del fenomen de representació, aleshores els anomenem operadors de funció temporal (v. apartat 4.2.3). Els operadors de funció temporal han d'operar tenint en compte les funcions de representació; així aquestes esdevenen paràmetres d'aquests operadors. Això no obstant, no podem definir un lligam genèric entre els operadors de funció temporal i les funcions de representació ja que aquestes darreres són una funció contínua al llarg del temps mentre que els operadors de funció temporal defineixen com manipular les mesures de la sèrie temporal.

Així doncs, el lligam entre els operadors de funció temporal i les funcions de representació és simbòlic: en els operadors de funció temporal, el paràmetre de representació és un nom que indica el concepte de representació en el que es basen els càlculs de l'operació. Per tant, per a cada parella d'operació de funció temporal i nom de representació cal definir quins càlculs s'han de dur a terme tot interpretant

el significat de la funció de representació associada, si bé només és necessari implementar l'operació d'interval temporal per a cada representació atès que les altres operacions es defineixen a partir d'aquesta.

A la definició de l'operació d'interval temporal (v. definició 4.39) hem deixat oberta la interpretació del mètode de representació. A continuació particularitzem l'interval temporal segons quatre representacions: DD, ZOHE, FOH i lineal.

Definició 4.55 (Interval temporal DD)**.** *Sigui S una sèrie temporal, $[s,t]$ un interval de temps i el mètode de representació* DD*. Definim l'interval temporal* DD*, $S[s,t]^{\text{DD}}$, com la sèrie temporal $S[s,t]^{\text{DD}} = S[s,t] \cup \{(s,0),(t,0)\}$.*

Definició 4.56 (Interval temporal ZOHE)**.** *Sigui S una sèrie temporal, $[s,t]$ un interval de temps i el mètode representació* ZOHE*. Definim l'interval temporal* ZOHE*, $S[s,t]^{\text{ZOHE}}$, com la sèrie temporal $S[s,t]^{\text{ZOHE}} = S(s,t] \cup \{m\}$ on $m = (t,v)$ i $v = V(\inf(S[t,+\infty]))$.*

Definició 4.57 (Interval temporal FOH)**.** *Sigui S una sèrie temporal, $[s,t]$ un interval de temps i el mètode de representació* FOH*. Definim l'interval temporal* FOH*, $S[s,t]^{\text{FOH}}$, com la sèrie temporal $S[s,t]^{\text{FOH}} = S[s,t] \cup \{m,n\}$ on $m = (s, S^{FOH}(s))$ i $n = (t, S^{\text{FOH}}(t))$.*

Definició 4.58 (Interval temporal lineal)**.** *Sigui S una sèrie temporal, $[s,t]$ un interval de temps i el mètode de representació lineal. Definim l'interval temporal lineal, $S[s,t]^{lineal}$, com la sèrie temporal $S[s,t]^{lineal} = \{m,n\}$ on $m = (s, S^{lineal}(s))$ i $n = (t, S^{lineal}(t))$.*

En conclusió, l'objectiu principal de les funcions de representació és estudiar exactament els grafs que es poden obtenir de la sèrie temporal per a posteriorment, agafant-ne el significat, implementar les operacions que siguin necessàries D'altra banda, les funcions de representació també es tindran en compte a l'hora de definir interpretacions pels agregadors d'atributs en el model de SGSTM (v. secció 5.3).

4.3.3. Patologies de les sèries temporals

Les sèries temporals poden no ser ideals a causa de problemes en la seva adquisició o del seu tractament. Tot i que segueixen sent sèries temporals, tenen unes propietats que poden resultar problemàtiques a l'hora d'operar-hi i que anomenem patologies. A continuació expliquem algunes patologies habituals.

Una primera patologia prové de problemes en el rellotge, com poden ser la precisió i l'exactitud en la mesura del temps. Kopetz [74, cap. 3] descriu en profunditat aquests problemes i conclou que els rellotges necessiten tècniques de sincronització, de les quals en descriu el funcionament.

Una segona patologia és la de les dades desconegudes. Quan no s'han capturat dades o quan s'han capturat erròniament aleshores s'han de tractar com a dades desconegudes; és a dir, s'ha de validar que les dades siguin correctes i en cas contrari rebutjar-les, tot i que cal tenir en compte que en alguns casos posteriorment es poden aplicar processos de reconstrucció per a aquestes dades errònies. Per a marcar dades com a desconegudes cal que el domini de valors inclogui un element específic (v. apartat 4.1.2) i el marcatge pot ocórrer en el mateix moment d'adquirir una mesura –és a dir quan no existeix valor capturat de la mesura i per tant es desconeix la dada (*missing data*)– o provenir d'un processament posterior de les dades– és a dir que les dades s'ignoren o es descarten (*censoring* o *truncation*). Així doncs, cal no confondre les dades desconegudes amb les dades no capturades: és a dir quan no s'han capturat dades perquè no hi havia intenció de capturar-les i per tant no es pot concloure com a fet cert ni que la dada és coneguda ni que és desconeguda. La patologia de dades desconegudes presenta diversitat en les causes, el domini de valors podria preveure un element per a marcar cada causa o usar el genèric:

- Els valors de les magnituds físiques estan limitats en un rang. Per tant, si s'adquireixen valors de fora del rang aquests no són vàlids. Per exemple, es captura un valor d'un sensor que és clarament fora d'uns límits raonables.
- En el moment de recol·lecció de dades pot aparèixer una mesura inexistent ja sigui perquè s'ha recollit un valor incomprensible o perquè no s'ha pogut recollir la mostra per manca de temps d'execució. Per exemple, s'intenta capturar una dada d'un sensor però aquest no respon o bé aquest respon amb un caràcter quan s'esperava un nombre.
- S'ha definit un temps de termini de l'adquisició de mesures; és a dir que si entre dues mesures hi ha una durada superior a un temps anomenat termini la mesura no és vàlida. Per exemple, es captura una dada d'un sensor però aquest no respon en un temps raonable.

Una tercera patologia és la gestió d'una quantitat enorme de dades. Les sèries temporal són els objectes que permeten gestionar de les dades recollides en els sistemes de monitoratge, que poden gestionar una gran quantitat de variables i en aquests casos pot haver-hi dificultats en l'operació i en la visualització de les sèries temporals, lentitud en executar les consultes o realització de càlculs innecessaris.

Una quarta patologia ocorre quan el període de mostreig no és regular, és a dir que les dades no es recullen de manera uniforme en el temps, però les aplicacions no ho preveuen o volen treballar amb dades a intervals regulars. A continuació aprofundim en aquest tema.

Regularitat de les sèries temporals

Per a determinar la regularitat d'una sèrie temporal es defineix un interval de temps $I_0 = [t, t+d]$, on t és un instant de temps i d una durada de temps, i els seus intervals múltiples $I_n = [t + nd\,,\ t + (n+1)d]$ per $n = 0, 1, 2 \ldots$. Aleshores, la regularitat de la sèrie temporal depèn de la situació dels temps de les seves mesures en aquests intervals de temps I_n.

Quan la situació temporal de les mesures prové del sistema d'adquisició de dades aleshores cal notar que, en l'àmbit de teoria del senyal, aquests intervals de temps I_n s'anomenen intervals de mostreig, d s'anomena període de mostreig i t s'anomena temps inicial del mostreig.

Una sèrie temporal és regular quan les mesures són equidistants en el temps, tal com ho anomena Hetland [57]. La regularitat d'una sèrie temporal és crítica en algunes operacions perquè hi ha algoritmes d'anàlisi de sèrie temporals que només es poden aplicar a sèries temporals regulars.

Definició 4.59 (Sèrie temporal regular)**.** *Sigui* $S = \{m_0, m_1, \ldots, m_{k-1}, m_k\}$ *la permutació d'una sèrie temporal en què* $T(m_0) < T(m_1) < \cdots < T(m_{k-1}) < T(m)$*, sigui* t *un instant de temps i sigui* d *una durada de temps. La sèrie temporal* S *és regular quan* $d = T(m_1) - T(m_0) = \cdots = T(m_k) - T(m_{k-1})$ *on definim* $t = T(m_0)$*.*

Si una sèrie temporal és regular, l'anomenem sèrie temporal regular de període d iniciada a t. Si t pot ser qualsevol instant de temps llavors simplement l'anomenem sèrie temporal regular de període d. Si notem l'àmbit de teoria del senyal, aquest període té relació amb el període de mostreig.

Així doncs, per contraposició definim que una sèrie temporal és *no regular* o *irregular* quan no és regular. A continuació distingim tres característiques que es poden definir per a les sèries temporals no regulars: temps real, ultramostreig i inframostreig.

L'adquisició de les mesures pot estar sotmesa a un sistema de control de temps real. Aquest sistema sol ser el que mana sobre el temps de mostreig ja que les periodicitats per a control són més elevades que per a altres necessitat, com per exemple la visualització de les dades per part de persones, les quals tenen un temps de resposta més lent [74, cap. 1]. El temps real causa que apareguin sèries temporals no regulars però amb unes certes característiques de periodicitat. Així, seguint el vocabulari de temps real, en les sèries temporals no regulars podem distingir els tres casos mencionats

Una sèrie temporal és de temps real quan a cada interval de mostreig hi ha una i només una mesura. A més, l'interval de mostreig pot estar fitar per una durada anomenada termini.

Definició 4.60 (Sèrie temporal de temps real)**.** *Sigui S una sèrie temporal, t un instant de temps, d una durada de temps i D una durada que indica termini. La sèrie temporal S és de temps real si i només si $D \leq d$ i la subsèrie de cada interval amb termini només conté una mesura* $\forall n \in \{0, \ldots, |S|-1\} : |S[t+nd, t+nd+D)| = 1$.

Si una sèrie temporal és de temps real, l'anomenem sèrie temporal mostrejada en temps real de període d iniciada a t i compliment del termini D. Si $D = d$, es pot anomenar que S és una sèrie temporal de temps real sense termini.

Pel que fa al temps inicial t, en el cas de les sèries temporals de temps reals pot estar situat entre $T(\min(S)) - \delta < t \leq T(\min(S))$. Així doncs, una sèrie temporal regular és també de temps real. En el cas que una sèrie temporal compleixi els requeriments de regular a excepció de $T(\min(S)) = t$, és a dir de la qual les mesures són equidistants però no té el temps inicial demanat, aleshores només és una sèrie temporal de temps real.

Seguint en l'àmbit de temps real, quan no es compleix que a cada interval de mostreig hi ha una i només una mesura pot succeir que hi hagi un, o més d'un, interval amb cap mesura o amb més d'una. Respectivament, ho anomenem interval inframostrejat i interval ultramostrejat.

Una sèrie temporal té intervals amb ultramostreig (*upsampling*) quan en alguns intervals de mostreig hi ha una mesura o més d'una.

Definició 4.61 (Sèrie temporal amb ultramostreig)**.** *Sigui S una sèrie temporal, t un instant de temps i d una durada de temps, els intervals amb ultramostreig de S són aquells en què la subsèrie corresponent conté més d'una mesura* $|S[t + nd, t + (n + 1)d)| > 1$ *on* $n \in \mathbb{N}$.

Una sèrie temporal té intervals amb inframostreig (*downsampling*) quan en alguns intervals de mostreig no hi ha cap mesura.

Definició 4.62 (Sèrie temporal amb inframostreig)**.** *Sigui S una sèrie temporal, t un instant de temps i d una durada de temps, els intervals amb inframostreig de S són aquells en què la subsèrie corresponent no conté cap mesura* $|S[t + nd, t + (n + 1)d)| = 0$ *on* $n \in \mathbb{N}$.

En resum, una sèrie temporal no regular es pot classificar segons si és de temps real i en el cas que no ho sigui es pot indicar si té intervals amb ultramostreig o si en té amb inframostreig. Aquests dos darrers casos són possibles alhora per una mateixa sèrie temporal però no en un mateix interval.

En el cas de dues sèries temporals, també es pot establir la relació de regularitat que tenen segons com són els vectors de temps.

Definició 4.63 (Sèries temporals regulars entre elles)**.** *Siguin S_1 i S_2 dues sèries temporals, S_1 i S_2 són regulars entre elles si i només si tenen el mateix vector de temps: $\Pi_t(S_1) = \Pi_t(S_2)$. D'aquesta manera també han de tenir el mateix cardinal $|S_1| = |S_2|$.*

Regularització de sèries temporals

Per a regularitzar una sèrie temporal, és a dir per a obtenir una sèrie temporal regular d'una no regular o per a canviar el període d'una regular, s'han d'utilitzar operadors per a generar noves mesures que compleixin amb les restriccions de temps regulars.

Un dels operadors que faciliten la feina de regularitzar són els de funció temporal, en aquests casos la tècnica de regularitzar es basa fortament en les funcions de representació (v. apartat 4.3.2). Així, per exemple, la selecció temporal (v. definició 4.40) permet regularitzar una sèrie temporal. Sigui S una sèrie temporal, t i d els paràmetres de temps desitjats de regularitat i $k \in \mathbb{N}$ una fita del rang. Una S regular es pot obtenir mitjançant $S[I]$ on $I = \{t + nd | n \in \mathbb{N} \wedge n \leq k\}$.

A continuació mostrem algunes tècniques simples per a regularitzar les sèries temporals de temps real, les d'ultramostreig i les d'inframostreig. Aquestes tenen determinats paràmetres de periodicitat que faciliten la conversió a regular, o dit d'una altra manera, es pot planificar des d'un inici el seu mostreig per a posteriorment associar-ho a un mètode de regularització adequat.

En primer lloc, de manera senzilla es pot assumir que una sèrie de temps real és regular amb un marge d'error tolerable ja que és causat per la impossibilitat d'efectuar els mostrejos exactament en el moment desitjat. Així en una sèrie temporal de temps real es poden canviar els instants de temps de cada mesura a aquells que s'ajusten al període de mostreig teòric; és a dir es pot assumir que cada mesura és vàlida per a tot l'interval de mostreig.

Sigui S una sèrie temporal de temps real de període d iniciada a t, una operació de funció temporal que regularitza de manera simple el temps real és la selecció temporal amb representació ZOHE, $S[I]^{\text{ZOHE}}$, amb els instants de temps $I = \{t + nd | n \in \mathbb{N}, t + nd < T(\sup(S))\} : .$

En segon lloc, de manera senzilla es pot assumir que una sèrie temporal amb ultramostreig és regular amb intervals en què s'ha adquirit més d'una mesura, els quals cal reduir-los a una de sola. El cas més simple és descartar les que es trobin més lluny de l'instant de temps de mostreig regular, per tant l'operació és la mateixa que en el cas anterior, $S[I]^{\text{ZOHE}}$.

Un altre cas, també simple, és el d'agregar les mesures dels intervals amb ultramostreig, per exemple amb la mitjana, el màxim, etc. Proposem un exemple pel cas de la mitjana.

Sigui S una sèrie temporal d'ultramostreig de període d iniciada a t i sigui I el conjunt desitjat d'instants de temps regulars. La regularització per agregació mitjana és mapa(R, f) on $R = \{(t, 0)|t \in I\}$ i $f(m) = (T(m), \text{mitjana\_v}\,(S[T(m), \text{seg}_R(T(m)))))$.

Aquesta regularització també es pot expressar mitjançant operadors de funció temporal. Sigui una funció de representació de mitjana a trossos de la qual definim l'operador d'interval temporal com $S[s,t]^{\text{mitjana}} = \{(s, v)\}$ on $v = \text{mitjana\_v}(S[s,t))$. L'operació temporal corresponent a la regularització per mitjana de l'ultramostreig és $S[I]^{\text{mitjana}}$.

En tercer lloc, per a les sèries temporals amb inframostreig es poden aplicar tècniques de farciment de forats; és a dir tècniques de reconstrucció. Un cas simple de farciment és utilitzar el valor del següent interval. Per tant l'operació també és la mateixa que en els casos simples anteriors $S[I]^{\text{ZOHE}}$.

En el cas de farcir forats inframostrejats normalment hi ha un durada vàlida per a fer el farciment, és a dir un temps de termini que indica a partir del qual els intervals no es podran farcir i les mesures s'hauran de considerar desconegudes. Aquest temps de termini és diferent del concepte de termini de temps real descrit anteriorment ja que el primer és una durada entre mesures i el segon és una durada amb relació al període regular. Oetiker [97] anomena *heartbeat* a aquest temps de termini entre mesures.

5. Model SGSTM

En aquest capítol es defineix un model dels sistemes de gestió de bases de dades per a sèries temporals multiresolució (SGSTM), el qual permet emmagatzemar sèries temporals de forma resumida i compacta. Aquest model s'estructura en uns objectes principal que són les *sèries temporals multiresolució*, les quals es defineixen com a conjunts de *subsèries resolució* formades per *discs* i *buffers*, i utilitza els objectes de sèries temporals i mesures descrits en el model de SGST. El model es dissenya en tres parts.

- Primer, es defineix el model d'estructura de les dades, és a dir, la forma que prenen els buffers, els discs, les subsèries resolució, les sèries temporals multiresolució i les bases de dades de sèries temporals multiresolució.
- Segon, es defineix el model d'operacions sobre les dades, és a dir, els operadors bàsics que permeten modelar el comportament i la manipulació de bases de dades multiresolució.
- Tercer, s'expliquen amb més detall les funcions específiques del model que permeten agregar diferents atributs de les sèries temporals per a obtenir les multiresolucions.

5.1. Model estructural de dades

Els objectes estructurals principals d'un SGSTM són els següents:

- Buffer
- Disc
- Subsèrie resolució
- Sèrie temporal multiresolució
- Esquema de multiresolució

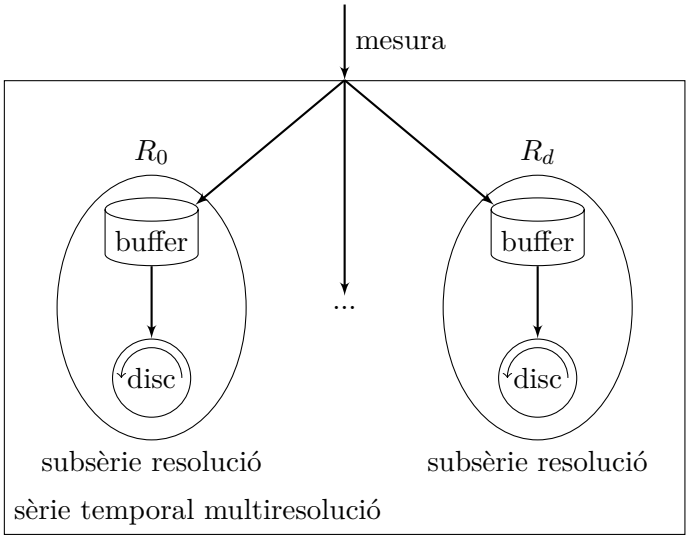


Figura 5.1.: Arquitectura d'una base de dades multiresolució

Definirem formalment cadascun d'aquests objectes en els apartats següents. Aquests objectes permeten definir l'arquitectura d'una base de dades multiresolució, la qual es pot veure gràficament a la figura 5.1. Una sèrie temporal multiresolució és una col·lecció de subsèries resolució, les quals acumulen temporalment mesures en un buffer a on són processades i finalment emmagatzemades en un disc. Aquest processament té per objectiu canviar els intervals de temps entre les mesures per tal de compactar la informació de les sèries temporals i emmagatzemar-la en forma de subsèries temporals amb diferents resolucions distribuïdes en els discs. Els discs tenen la mida limitada i només poden contenir un nombre fixat de mesures.

Resumint, una base de dades multiresolució és una solució d'emmagatzematge per a sèries temporals en què la informació de cadascuna es distribueix mitjançant diferents resolucions temporals. Així, una base de dades multiresolució és un contenidor de sèries temporals multiresolució on cadascuna és una subsèrie resolució formada per una relació d'un buffer amb un disc. Cada sèrie temporal multiresolució té un esquema de multiresolució que indica de quina manera s'ha de resumir la informació per calcular les resolucions.

En aquesta secció es defineixen els conceptes referents a l'estructura del model i al final s'ofereixen alguns exemples amb valors concrets. Aquesta estructura, però, requereix uns operadors específics per a emmagatzemar-hi i consolidar-hi les mesures, els quals es defineixen a l'apartat 5.2.1. També requereix unes funcions per a agregar els atributs d'una sèrie temporal, les quals es defineixen a la secció 5.3.

5.1.1. Buffer

Un buffer és un contenidor d'una sèrie temporal que s'ocupa d'emmagatzemar les mesures pendents de ser consolidades i de consolidar-les en funció d'unes característiques definides prèviament. Les mesures es poden eliminar un cop s'han consolidat.

La consolidació d'un buffer té com a objectiu regularitzar la sèrie temporal a un període de mostreig constant (v. definició 4.59 i secció 4.3.3). Així, els paràmetres d'un buffer són: el període temporal de consolidació, que anomenen pas de consolidació; l'instant inicial de consolidació, que s'expressa com el darrer instant de temps de consolidació; i la funció que extreu les característiques de la consolidació, el càlcul de la qual es delega al que anomenem funció d'agregació d'atributs.

Definició 5.1 (Buffer)**.** *Sigui S_B una sèrie temporal pendent de ser consolidada, τ el darrer instant de temps de consolidació, δ una durada de temps que indica el pas de consolidació i f una funció d'agregació d'atributs. Definim un* buffer *B com el tuple $B = (S_B, \tau, \delta, f)$.*

La consolidació d'una sèrie temporal s'inicia en un instant de temps concret, τ, i ocorre a cada pas de consolidació, δ. Amb la finalitat d'establir els intervals de consolidació de la sèrie temporal, es defineix un buffer inicial.

Definició 5.2 (Buffer buit)**.** *Definim buffer inicial o buffer buit com el buffer $(\emptyset, t_0, \delta, f)$, el qual conté una sèrie temporal buida, un instant de temps inicial de consolidació $t_0 \in \mathcal{T}$, el pas de consolidació δ i una funció d'agregació d'atributs f.*

A partir d'un buffer es poden conèixer tots els instants de temps de consolidació, els quals seran $\tau_n = t_0 + n\delta$ per tot $n \in \mathbb{N}$. Aquests instants de temps de consolidació també defineixen els intervals de temps de consolidació del buffer de la forma $[\tau_n, \tau_n + \delta]$. La consolidació de la sèrie temporal S_B d'un buffer en un interval de temps $[\tau, \tau + \delta]$ dóna com a resultat una mesura m calculada a partir de la funció d'agregació d'atributs $m = f(S_B, \tau, \delta)$. Més endavant a la secció 5.3 detallem aquest concepte de funció d'agregació d'atributs.

A la figura 5.2 es mostra un exemple de sèrie temporal les mesures de la qual es mostren amb punts vermells, en línies blaves verticals s'indiquen els intervals de consolidació d'un buffer sobre aquesta sèrie temporal, i les mesures consolidades es mostren amb punts verds. Sigui t_0 l'instant inicial de consolidació, el primer valor que pren τ és $\tau_0 = t_0$, el següent valor que pren és $\tau_1 = \tau_0 + \delta$, i així succesivament. Per tant el primer interval de consolidació és $[\tau_0, \tau_1]$. Quan es consolida el buffer en aquest aquest interval, el darrer instant de consolidació esdevé τ_1 i es calcula una mesura $m_1 = f(S, \tau_0, \delta)$; la qual per facilitar la comprensió podem assumir de moment que és el resultat de calcular una agregació sobre les mesures en l'interval $[\tau_0, \tau_1]$. De la mateixa manera es calculen les mesures consolidades m_2 i m_3.

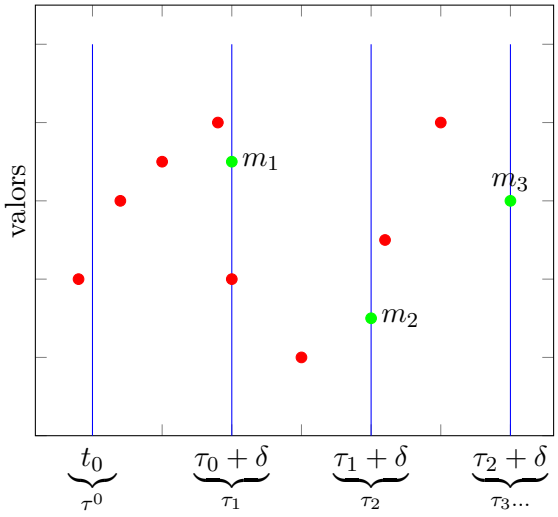


Figura 5.2.: Intervals de consolidació d'un buffer

5.1.2. Disc

Un disc és un contenidor d'una sèrie temporal regular amb un nombre acotat de mesures. En arribar al nombre màxim de mesures permeses, cada cop que s'afegeix una mesura nova s'elimina la mesura mínima de la sèrie temporal. Així doncs, un disc és semblant a una cua *First In First Out*, en què el primer d'arribar és el primer de sortir.

Definició 5.3 (Disc)**.** *Sigui S_D una sèrie temporal regular i $\kappa \in \mathbb{N}$ el cardinal màxim que pot prendre, $|S_D| \leq \kappa$. Definim un* disc *D com el tuple $D = (S_D, \kappa)$.*

A l'inici, un disc no conté mesures però cal que estigui caracteritzat pel cardinal màxim. Amb aquesta finalitat es defineix un disc inicial.

Definició 5.4 (Disc buit)**.** *Definim disc inicial o disc buit com el disc $(\emptyset, \kappa)$, el qual conté una sèrie temporal buida i el cardinal màxim κ que podrà pendre.*

5.1.3. Subsèrie resolució

Una subsèrie resolució és una parella de disc i buffer. En el buffer hi ha la part d'una sèrie temporal a regularitzar i en el disc hi ha l'altra part ja regularitzada, amb un nombre afitat de mesures. A l'acció de regularitzar l'anomenem consolidar en coherència amb el concepte descrit pels buffers.

Definició 5.5 (Subsèrie resolució)**.** *Sigui B un buffer i D un disc. Definim una* subsèrie resolució *R com el tuple $R = (B, D)$.*

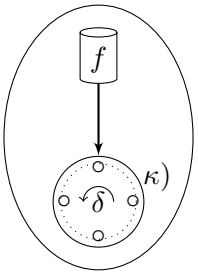


Figura 5.3.: Paràmetres d'una subsèrie temporal resolució

La definició 5.2 de buffer buit i la definició 5.4 de disc buit indueixen a una definició de subsèrie resolució buida.

Definició 5.6 (Subsèrie resolució buida)**.** *Sigui B un buffer buit i D un disc buit. Definim subsèrie resolució buida com la subsèrie resolució (B, D).*

Les subsèries resolució es consoliden seguint els criteris del seu buffer i emmagatzemant al seu disc la mesura que resulta de la consolidació. Així, els paràmetres del buffer i del disc d'una subsèrie resolució $(\tau, \delta, f, \kappa)$ caracteritzen la resolució de la subsèrie temporal S_D que finalment queda emmagatzemada. A la figura 5.3 es mostra un esquema semblant a la figura 5.1 on es veu una subsèrie resolució amb la relació d'aquests paràmetres entre el buffer i el disc, llevat de τ que és difícil de representar.

5.1.4. Sèrie temporal multiresolució

Una sèrie temporal multiresolució és un conjunt de subsèries resolució que comparteixen l'entrada de mesures, les quals provenen d'una mateixa sèrie temporal. Aquesta sèrie temporal queda regularitzada i distribuïda en les diferents subsèries resolució amb paràmetres diferents, tal com s'ha vist a la figura 5.1

Definició 5.7 (Sèrie temporal multiresolució)**.** *Definim una* sèrie temporal multiresolució *M com el conjunt de subsèries resolució $M = \{R_0, \ldots, R_k\}$.*

A partir de la definició 5.6 de subsèrie resolució buida és defineix la sèrie temporal multiresolució buida.

Definició 5.8 (Sèrie temporal multiresolució buida)**.** *Definim sèrie temporal multiresolució buida o inicial com el conjunt de subsèries resolució buides $\{R_0, \ldots, R_k\}$.*

Normalment, en una sèrie temporal multiresolució no hi ha dues subsèries resolució amb la mateixa informació. És a dir, no hi dues subsèries resolució els buffers de les quals tinguin alhora el mateix pas de consolidació i funció d'agregació d'atributs: $|\{(\delta, f) | (B, D) \in M, B = (S_B, \tau, \delta, f)\}| = |M|$.

Relació sèrie temporal multiresolució

Una sèrie temporal multiresolució s'expressa com un conjunt i com a tal és susceptible d'aplicar-hi els conceptes del model dels SGBDR (v. § 2.2), a continuació expressem la forma de sèrie temporal multiresolució seguint també el concepte de relació.

Una sèrie temporal multiresolució és una relació de buffers i discs. A cada parella buffer-disc l'anomenem subsèrie resolució. Així doncs, una sèrie temporal multiresolució és un conjunt de subsèries resolució. Com a conjunt de subsèries resolució, una sèrie temporal multiresolució s'observa com una relació de grau sis en què la capçalera conté els atributs

- sèrie temporal del buffer (S_B),
- sèrie temporal del disc (S_D),
- darrer instant de consolidació (τ),
- pas de consolidació (δ),
- màxim cardinal del disc (κ),
- i funció d'agregació d'atributs (f).

Com ja s'ha comentat, una restricció habitual és que δ i f no siguin repetits, és a dir que és una restricció que indica que $\{\delta, f\}$ són els attributs clau. El predicat corresponent a les sèries temporals multiresolució és similar a: «La resolució amb pas de consolidació δ i funció d'agregació f, de la qual se n'emmagatzemarà com a màxim κ mesures, s'ha consolidat amb la sèrie temporal S_D per darrer cop a l'instant τ i té pendent per consolidar la sèrie temporal S_B».

Així doncs observada com a relació, i tal com s'ha fet per a la forma canònica de les sèries temporals (v. definició 4.9), podem escriure la forma canònica d'una sèrie temporal multiresolució.

Definició 5.9 (Forma canònica)**.** *Sigui $M = \{R_0, \ldots, R_k\}$ una sèrie temporal multiresolució on cada subsèrie resolució té la forma $R = (B, D), B = (S_B, \tau, \delta, f)$, $D = (S_D, \kappa)$ i té domini de sèrie temporal per les sèries temporals S_B i S_D, de $\bar{\mathbb{R}}$ pels temps τ i δ, de $\mathbb{N}$ pel cardinal k, i de funció per a la funció d'agregació d'atributs. En la forma canònica s'escriu com*

$$
\begin{aligned}
M = (\{&S_B, S_D, \tau, \delta, \kappa, f\}, \{ \\
&\{(S_B, S_{B_0}), (S_D, S_{D_0}), (\tau, \tau_0), (\delta, \delta_0), (\kappa, \kappa_0), (f, f_0)\}, \\
&\ldots, \\
&\{(S_B, S_{B_k}), (S_D, S_{D_k}), (\tau, \tau_k), (\delta, \delta_k), (\kappa, \kappa_k), (f, f_k)\} \\
\})&
\end{aligned}
$$

De la mateixa manera que per les sèries temporals, també escrivim una sèrie temporal multiresolució de manera simplificada com el conjunt de tuples $M = \{(S_{B_0}, S_{D_0}, \tau_0, \delta_0, \kappa_0, f_0), \ldots, (S_{B_k}, S_{D_k}, \tau_k, \delta_k, \kappa_k, f_k)\}$, la qual es correspon amb la forma expressada inicialment a la definició 5.7 amb els tuples de B i de D units. També com amb les sèries temporals, les sèries temporals multiresolució es poden visualitzar com a taules, cosa que mostrem en exemples posteriors.

5.1.5. Esquema de multiresolució

Una sèrie temporal multiresolució té paràmetres que s'han de configurar en el seu estat inicial; la quantitat de subsèries resolucions i els paràmetres de cada una: pas de consolidació, funció d'agregació d'atributs, darrer instant de consolidació i cardinal màxim. Anomenem esquema de multiresolució a les configuracions possibles d'aquests paràmetres.

En la forma canònica de les sèries temporals multiresolució es pot observar més clarament que els atributs $\{\delta, \tau, f, \kappa\}$ són els paràmetres configurables al conjunt de tuples dels quals anomenem esquema de multiresolució.

Definició 5.10 (Esquema de multiresolució)**.** *Sigui $M = \{R_0, \ldots, R_k\}$ una sèrie temporal multiresolució, el seu esquema de multiresolució E és la projecció dels atributs que constitueixen els paràmetres configurables: $E = \Pi_{\{\delta,\tau,f,\kappa\}}(M)$. Així, l'esquema de multiresolució de M es pot expressar com la relació $E = \{(\delta_0, \tau_0, f_0, \kappa_0), \ldots, (\delta_k, \tau_k, f_k, \kappa_k)\}$, és a dir una relació amb la capçalera $\{\delta, \tau, f, \kappa\}$.*

Per a cada sèrie temporal multiresolució es pot estudiar i manipular l'esquema de multiresolució, la qual cosa mostrem amb més detall a l'apartat 5.2.2.

5.1.6. Exemples

Exemple 5.1 (Sèrie temporal multiresolució)**.** Sèrie temporal multiresolució $M_1 = \{R_1, R_2\}$ que té dues subsèries resolució amb els paràmetres següents:

- La subsèrie resolució R_1 té un pas de consolidació de 5 unitats de temps, una mida màxima de 4 mesures i una funció de consolidació de 'mitjana' de les mesures.
- La subsèrie resolució R_2 té un pas de consolidació de 10 unitats de temps, una mida màxima de 3 mesures i una funció de consolidació de 'mitjana' de les mesures.

L'arquitectura de la base de dades que conté aquesta sèrie temporal multiresolució es pot veure a la figura 5.4. L'esquema de multiresolució que correspon als instants de consolidació, des de 0 fins a 30, és el següent:

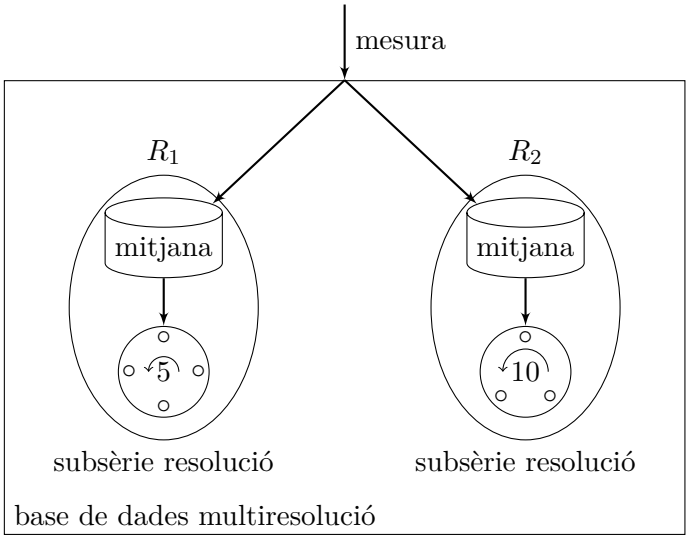


Figura 5.4.: Arquitectura de la base de dades multiresolució particular per l'exemple 5.1

- La subsèrie resolució R_1 serà consolidada en els instants 5, 10, 15, 20, 25 i 30.
- La subsèrie resolució R_2 serà consolidada en els instants 10, 20 i 30.

Iniciem la base de dades a l'instant de temps 0, instant en el qual la sèrie temporal multiresolució és $M_1^0 = \{(\{\}, \{\}, 0, 5, 4, \text{mitjana}), (\{\}, \{\}, 0, 10, 3, \text{mitjana})\}$; és a dir amb les sèries temporals buides i els darrers instants de consolidació iniciats a 0.

A continuació, afegim a la sèrie temporal multiresolució les mesures de la sèrie temporal $S_1 = \{(1,0), (5,0), (8,0), (10,0), (14,0), (19,0), (22,0), (26,0), (29,0)\}$. Tots els valors valen zero per tal de centrar la comprensió de l'exemple en l'estructura de temps de consolidació; pel que fa a exemples d'agregació de valors es poden veure amb més detall a la secció 5.3.

Si consolidem la sèrie temporal multiresolució cada cop que sigui consolidable, és a dir en els instants que marca l'esquema de multiresolució, a l'instant 29 després d'haver inserit la darrera mesura la sèrie temporal multiresolució és $M_1^{29} = \{(\{(26,0), (29,0)\}, \{(10,0), (15,0), (20,0), (25,0)\}, 25, 5, 4, \text{mitjana}), (\{(22,0), (26,0), (29,0)\}, \{(10,0), (20,0)\}, 20, 10, 3, \text{mitjana})\}$. Aquesta sèrie temporal multiresolució es mostra a la figura 5.5 en forma de taula.

Es pot observar que als buffers hi ha emmagatzemades les mesures pendents de consolidar per a cada subsèrie i als discs les darreres mesures consolidades:

- Per a la subsèrie resolució R_1 hi ha pendent de consolidar l'interval de temps $[25, 30]$ i al disc hi ha emmagatzemades les 4 mesures màximes permeses; és a dir que la que s'havia consolidat a l'instant 5 ja s'ha perdut.

M_1^{29}

S_B	S_D	τ	δ	κ	f
t | v 26 | 0 29 | 0	t | v 10 | 0 15 | 0 20 | 0 25 | 0	25	5	4	mitjana
t | v 22 | 0 26 | 0 29 | 0	t | v 10 | 0 20 | 0	20	10	3	mitjana

Figura 5.5.: Taula d'una sèrie temporal multiresolució a l'instant 29

- Per a la subsèrie resolució R_2 hi ha pendent de consolidar l'interval de temps $[20, 30]$ i al disc hi ha emmagatzemades 2 mesures. El disc encara no ha arribat al cardinal màxim $\kappa = 3$ a causa que la base de dades s'ha iniciat a l'instant 0 i la primera consolidació d'aquesta subsèrie ha estat a l'instant 10.

Exemple 5.2 (Sèrie temporal multiresolució amb vistes)**.** En el model SGBDR molt sovint s'utilitzen vistes per a agrupar informació de vàries relacions, per a mostrar-ne una part, etc. Una vista és una variable relació virtual derivada d'una expressió relacional [27]. En aquest exemple mostrem la mateixa sèrie temporal multiresolució de l'exemple 5.1 però organitzada amb forma de vistes relacionals.

Sigui S un atribut amb domini de sèrie temporal, siguin S', S'_B, S'_D uns atributs amb domini de noms, i siguin τ, δ, κ, f els atributs de l'esquema de multiresolució amb els dominis corresponents. Siguin una relació de noms i sèries temporals $L = ((S', S), \{(S_{B1}, \{(26, 0), (29, 0)\}), (S_{B2}, \{(22, 0), (26, 0), (29, 0)\}), (S_{D1}, \{(10, 0), (15, 0), (20, 0), (25, 0)\}), (S_{D2}, \{(10, 0), (20, 0)\})\})$ i una sèrie temporal multiresolució amb noms com a domini dels atributs de sèries temporals $M = ((S'_B, S'_D, \tau, \delta, \kappa, f), \{(S_{B1}, S_{D1}, 25, 5, 4, \text{mitjana}), (S_{B2}, S_{D2}, 20, 10, 3, \text{mitjana})\})$, les quals es mostren a la figura 5.6. Definim la vista de la sèrie temporal multiresolució com

$$\begin{aligned}\text{vista } M_2 =& \Pi_{\text{tots llevat } \{S'_B, S'_D\}}(M \bowtie (\rho_{\{S'/S'_B, S/S_B\}} L) \\ & \bowtie (\rho_{\{S'/S'_D, S/S_D\}} L))\end{aligned}$$

aplicant les operacions relacionals de reanomena, junció i projecció.

D'aquesta manera M_2 té els mateixos valors que la M_1 definida a l'exemple anterior és a dir que es visualitza amb la mateixa taula que la de la figura 5.5; observant només el resultat de M_2 no es pot distingir que és una vista. Així doncs, les vistes permeten organitzar una sèrie temporal multiresolució de forma més còmoda i a més, tal com descriu Date [27], mantenint que totes les operacions i propietats que són d'aplicació a les relacions ho són també a les seves vistes.

M

S'_B	S'_D	τ	δ	κ	f
S_{B1}	S_{D1}	25	5	4	mitjana
S_{B2}	S_{D2}	20	10	3	mitjana

L

S'	S	
	t	v
S_{B1}	26	0
	29	0
S_{B2}	22	0
	26	0
	29	0
S_{D1}	10	0
	15	0
	20	0
	25	0
S_{D2}	10	0
	20	0

Figura 5.6.: Taula d'una sèrie temporal multiresolució amb vistes relacionals

Exemple 5.3 (Sèrie temporal multiresolució amb desfasaments)**.** A l'exemple 5.1 s'ha mostrat una sèrie temporal multiresolució en què la consolidació de les dues subsèries obeeix a la mateixa funció d'agregador d'atributs. En aquest exemple treballem amb els mateixos valors que a l'exemple 5.1 però ara canviem la funció de la segona subsèrie resolució per un agregador amb desfasament; és a dir que cada cop que consolida retorna una mesura amb un retard d'una certa durada. Aquest nou agregador que anomenem *mitjanad5* també fa la mitjana però amb un desfasament de 5 unitat de temps, a l'apartat 5.2.2 es defineix amb més precisió aquest concepte de desfasament.

Seguint el mateix procediment que a l'exemple 5.1, a l'instant 29 després d'haver inserit la darrera mesura la sèrie temporal multiresolució és $M_3^{29} = \{(\{(26,0),(29,0)\}, \{(10,0),(15,0),(20,0),(25,0)\}, 25, 5, 4, \text{mitjana}), (\{(19,0),(22,0),(26,0),(29,0)\}, \{(5, 0),(15,0)\}, 20, 10, 3, \text{mitjanad5})\}$. Aquesta sèrie temporal multiresolució es mostra a la figura 5.7 en forma de taula.

Així doncs, mentre que l'esquema de multiresolució segueix sent el mateix pel que fa als instants de consolidació, els instants de temps de la sèrie temporal emmagatzemada a la subsèrie resolució R_2 tenen un retard de 5 unitats de temps. Per una banda, es pot observar a la S_{B2} que el buffer ara és 5 unitats de temps més gran i emmagatzema mesures de l'interval $[15, 30]$. Per altra banda, es pot observar a la S_{D2} que els instants emmagatzemats són 5 i 15 corresponents als instants de consolidació 10 i 20. Pel que fa a la resta de valors, no han variat respecte de l'exemple 5.1.

M_3^{29}

S_B	S_D	τ	δ	κ	f
t | v 26 | 0 29 | 0	t | v 10 | 0 15 | 0 20 | 0 25 | 0	25	5	4	mitjana
t | v 19 | 0 22 | 0 26 | 0 29 | 0	t | v 5 | 0 15 | 0	20	10	3	mitjanad5

Figura 5.7.: Taula d'una sèrie temporal multiresolució amb desfasaments

5.2. Model d'operacions

En aquesta secció es defineixen els operadors que permeten modelar el comportament i la manipulació de les dades en el model de SGSTM.

Per a treballar amb les sèries temporals multiresolució s'utilitzen els conceptes descrits al model d'operacions de SGST. El model de SGSTM es defineix a partir del model de SGST i per tant les operacions dels SGSTM també hi estan basades. Tot i així cal tenir en compte dues particularitats.

Per una banda, el model de SGSTM té una estructura específica que requereix ser manipulada coherentment. Així, es defineixen operadors que saben treballar amb aquesta estructura agrupats en dos grups. El primer grup són els operadors requerits pel model estructural; operadors que són inseparables de l'estructura i són utilitzats en el procés d'emmagatzemar les mesures. El segon grup són els operadors necessaris per a manipular l'estructura; és a dir operadors que permeten fer canvis en l'esquema de la base de dades o consultar paràmetres de l'esquema actual.

Per altra banda, el model de SGSTM treballa amb sèries temporals multiresolució. Així, es defineixen operadors que permeten extreure les sèries temporals emmagatzemades en aquestes bases de dades amb l'objectiu d'aplicar-hi posteriorment els operadors dels SGST.

En el disseny del model d'operacions següent es distingeixen tres grups d'operadors segons els casos anteriors:

- Estructurals: operadors requerits pel model estructural.
- Manipulació de l'esquema: operadors per a manipular l'esquema de multiresolució.
- Consultes: operadors per a extreure les sèries temporals emmagatzemades.

5.2.1. Estructurals

En el model estructural de SGSTM hem definit les sèries temporals multiresolució com un conjunt de subsèries resolució a les quals es van afegint mesures compactant-les i consolidant-les. En aquest apartat definim els operadors que permeten inserir mesures noves i consolidar-les al seu lloc corresponent en l'estructura.

A continuació es descriuen els operadors associats a cada objecte del model de SGSTM.

Buffer

Els buffers reben les noves mesures i les consoliden a cada instant de consolidació. Així, tenen dos operadors associats: un per afegir noves mesures al buffer i un altre per consolidar-les.

L'operació d'afegir una mesura al buffer consisteix en afegir-la a la sèrie temporal pendent de consolidar.

Definició 5.11 (Afegeix mesura al buffer)**.** *Sigui $B = (S, \tau, \delta, f)$ un buffer i m una mesura, la inserció de la mesura al buffer, notada* afegeixB(B, m)*, retorna un nou buffer amb la mesura afegida a la sèrie temporal del buffer:* afegeixB$(B, m) = (S', \tau, \delta, f)$ *on* $S' = S \cup \{m\}$.

L'operació de consolidació d'un buffer consisteix en compactar les mesures segons els intervals de consolidació i la funció d'agregació i a suprimir la part ja consolidada de la sèrie temporal. Així doncs, la consolidació d'un buffer per cada interval de temps $[\tau, \tau + \delta]$ dóna com a resultat una mesura calculada en funció de l'agregador d'atributs. També en la consolidació es pot calcular un nou buffer descartant les mesures ja consolidades.

Definició 5.12 (Consolida el buffer)**.** *Sigui $B = (S, \tau, \delta, f)$ un buffer, la consolidació del buffer, notada* consolidaB(B)*, retorna un nou buffer amb el nou instant de consolidació i una nova mesura* consolidaB$(B) = (B', m')$ *on* $B' = (S', \tau + \delta, \delta, f)$ *i* $m' = f(S, \tau, \delta)$*. En el nou buffer, S' és el resultat d'eliminar les dades històriques que no es necessiten més.*

Sobre eliminar les dades històriques*: En el model teòric es pot donar $S' = S$. Això no obstant, a les implementacions normalment caldrà eliminar les dades ja no necessàries. Es poden eliminar, per exemple, retornant la sèrie temporal $S' = S[\tau + \delta, +\infty]$.*

De manera simplificada, hem definit que cada consolidació només s'aplica a l'interval de consolidació actual; així la consolidació total del buffer és l'aplicació successiva de l'operació de consolidació.

Aquesta consolidació successiva requereix que les mesures s'insereixen al buffer ordenades en el temps, sinó un cop duta a terme la consolidació les mesures inserides desordenades poden no ser tingudes en compte. Si es duu a terme aquesta inserció ordenada, aleshores un buffer té l'estat de consolidable quan el temps d'una mesura de la sèrie temporal és més gran que el següent instant de temps de consolidació del buffer.

Definició 5.13 (Buffer consolidable)**.** *Sigui $B = (S, \tau, \delta, f)$ un buffer i $m = \sup(S)$ la mesura suprema de la sèrie temporal del buffer. Definim que B és consolidable si i només si $T(m) \geq \tau + \delta$.*

Disc

Els discs reben les mesures consolidades per a emmagatzemar-les de forma afitada. Així, tenen un operador associat que afegeix noves mesures al disc i manté sota control el cardinal màxim

L'operació d'afegir una mesura al disc consisteix en afegir-la a la sèrie temporal. Quan se supera el cardinal permès, aleshores es descarten algunes mesures.

Definició 5.14 (Afegeix mesura al disc)**.** *Sigui $D = (S, \kappa)$ un disc i m una mesura, la inserció de la mesura al disc, notada* afegeixD(D, m)*, retorna un nou disc amb la mesura afegida a la sèrie temporal, la qual manté el cardinal màxim:* afegeixD$(D, m) = (S', \kappa)$ *on*

$$S' = \begin{cases} S \cup \{m\} & si\ |S| < \kappa \\ (S - \{\min(S)\}) \cup \{m\} & altrament \end{cases}$$

Subsèrie resolució

Les subsèries resolució són l'aparellament d'un buffer amb un disc. Així tenen dos operadors associats: un per afegir una mesura al buffer i un altre per consolidar el buffer i afegir la mesura resultant al disc

L'operació d'afegir una mesura a la subsèrie resolució consisteix en afegir-la al buffer.

Definició 5.15 (Afegeix mesura a la subsèrie resolució)**.** *Sigui $R = (B, D)$ una subsèrie resolució i m una mesura, la inserció de la mesura a la subsèrie resolució, notada* afegeixR(R, m)*, retorna una nova subsèrie resolució amb la mesura afegida al buffer:* afegeixR$(R, m) = (B', D)$ *on* $B' =$ afegeixB(B, m)*.*

L'operació de consolidar una subsèrie resolució consisteix en calcular una mesura de consolidació del buffer, en l'interval de consolidació actual, i desar-la al disc.

Definició 5.16 (Consolida la subsèrie resolució)**.** *Sigui $R = (B, D)$ una subsèrie resolució, la consolidació de la subsèrie resolució, notada* consolidaR(R)*, retorna una nova subsèrie resolució $R' = (B', D')$ on $(B', m') =$* consolidaB(B) *és la consolidació del buffer i $D' =$* afegeixD(D, m') *és la inserció la mesura consolidada al buffer.*

Una subsèrie resolució és consolidable quan ho és el seu buffer.

Definició 5.17 (Subsèrie resolució consolidable)**.** *Sigui $R = (B, D)$ una subsèrie resolució, definim que R és consolidable si i només si B és consolidable*

Sèrie temporal multiresolució

Les sèries temporals multiresolució són un conjunt de subsèries resolució. Així tenen dos operadors per a treballar globalment amb totes les subèries que contingui: un per a afegir una mesura a cada subsèrie i un altre per a consolidar cadascuna de les subsèries.

L'operació d'afegir una mesura a la sèrie temporal multiresolució consisteix en afegir-la a cadascuna de les subsèries resolució.

Definició 5.18 (Afegeix mesura a la sèrie temporal multiresolució)**.** *Sigui* $M = \{R_0, \ldots, R_k\}$ *una sèrie temporal multiresolució i* m *una mesura, la inserció de la mesura a la sèrie temporal multiresolució, notada* $\text{afegeixM}(M, m)$*, retorna una nova sèrie temporal multiresolució amb la mesura afegida a cada subsèrie resolució:* $\text{afegeixM}(M, m) = \{R'_0, \ldots, R'_k\}$ *on* $R'_i = \text{afegeixR}(R_i, m)$.

L'operació de consolidar una sèrie temporal multiresolució consisteix en consolidar cadascuna de les subsèries resolució que siguin consolidables.

Definició 5.19 (Consolida la sèrie temporal multiresolució)**.** *Sigui* $M = \{R_0, \ldots, R_k\}$ *una sèrie temporal multiresolució, la consolidació de la sèrie temporal multiresolució, notada* $\text{consolidaM}(M)$*, retorna una nova sèrie temporal multiresolució que consolida les subsèries resolució consolidables:* $\text{consolidaM}(M) = \{R'_0, \ldots, R'_k\}$ *on*

$$R'_i = \begin{cases} \text{consolidaR}(R_i) & \textit{si } R_i \textit{ és consolidable} \\ R_i & \textit{altrament} \end{cases}$$

Una sèrie temporal multiresolució és un conjunt de subsèries resolució. Per a tractar amb tots i cadascun dels elements del conjunt és útil tenir l'operació de mapatge, de manera similar a com s'ha definit per les sèries temporals.

Definició 5.20 (Mapa d'una sèrie temporal multiresolució)**.** *Sigui* $M = \{R_0, \ldots, R_k\}$ *una sèrie temporal multiresolució en què* $\text{dom}(R)$ *és el domini de les subsèries resolució i sigui* $g : \text{dom}(R) \to \text{dom}(R)$ *una funció sobre una subsèrie resolució que retorna una nova subsèrie resolució. El mapa de* g *a* M*, notat* $\text{mapa}(M, g)$*, retorna una nova sèrie temporal multiresolució resultant d'aplicar la funció a cada subsèrie resolució:* $\text{mapa}(M, g) = \{g(R_0), \ldots, g(R_k)\}$.

5.2.2. Manipulació de l'esquema

El model de SGSTM associa a cada sèrie temporal un esquema de multiresolució. En aquest apartat definim els operadors que permeten consultar i manipular aquest esquema de multiresolució de forma coherent amb el model de SGSTM. Segons s'ha descrit a la definició 5.10, l'esquema de multiresolució de cada sèrie temporal

multiresolució consisteix en el nombre de subsèries resolució i els quatre paràmetres de cadascuna: el darrer instant de consolidació (τ), el pas de consolidació (δ), el cardinal màxim (κ) i la funció d'agregació d'atributs (f). Així doncs, quan es manipula una base de dades multiresolució cal conservar o tractar adequadament aquests esquemes de multiresolució.

A continuació es descriuen operadors per a poder estudiar aquest esquema, operadors per a canviar-lo i operadors per a unir o ajuntar dos esquemes. Una sèrie temporal multiresolució és un conjunt de subsèries resolució i per tant podem observar també l'esquema de multiresolució com a conjunt de l'esquema de cada subsèrie. Així, per simplicitat, definim la majoria dels operadors de manipulació sobre les subsèries resolució, tot i que es poden estendre fàcilment a una sèrie temporal multiresolució mitjançant l'operació de mapatge (v. definició 5.20), on la funció de mapatge correspon a l'operador d'esquema que es vol aplicar a totes i cadascuna de les subsèries resolució.

Propietats de l'esquema

La configuració dels paràmetres de l'esquema de multiresolució infereix diverses propietats a les sèries temporals multiresolució. A continuació definim algunes de les propietats que es poden estudiar a partir d'un esquema de multiresolució.

A partir d'un esquema de multiresolució es pot dibuixar la situació relativa en el temps que prendran les mesures. A aquest dibuix de l'esquema de multiresolució l'anomenem cronograma. Per exemple, sigui la sèrie temporal multiresolució $M = \{R_1, R_2\}$ de l'exemple 5.3 dibuixem el seu cronograma a la figura 5.8 per a l'instant de temps 30, just abans de la seva consolidació. Les subsèries resolució de M tenen els paràmetres $\delta_1 = 5$, $\delta_2 = 10$, $\kappa_1 = 4$, $\kappa_2 = 3$, $f_1 = \text{mitjana}$ i $f_2 = \text{mitjanad5}$ on la funció mitjanad5, ja utilitzada a l'exemple 5.3, té un desfasament de 5 unitats de temps. Tot seguit definim els conceptes que apareixen al cronograma.

Els paràmetres κ, δ i f d'una subsèrie resolució són fixats per l'esquema, mentre que el paràmetre τ és fixat a un valor inicial i va essent canviat per l'operació de consolidació. Així doncs, les propietats que impliquin a τ dependran de l'instant temporal en què es faci la consulta i les que no l'impliquin seran fixes per a cada esquema.

Una propietat que observem en el cronograma és el lapse temporal d'una subsèrie resolució, és a dir la mida temporal que ocupa la sèrie temporal emmagatzemada en el disc.

Definició 5.21 (Lapse de la subsèrie resolució)**.** *Sigui $R = (S_B, S_D, \tau, \delta, \kappa, f)$ una subsèrie resolució, el lapse* $\text{lapseR}(R)$ *és una durada de temps que indica la mida de l'interval que ocupa el disc:* $\text{lapseR}(R) = \kappa\delta$.

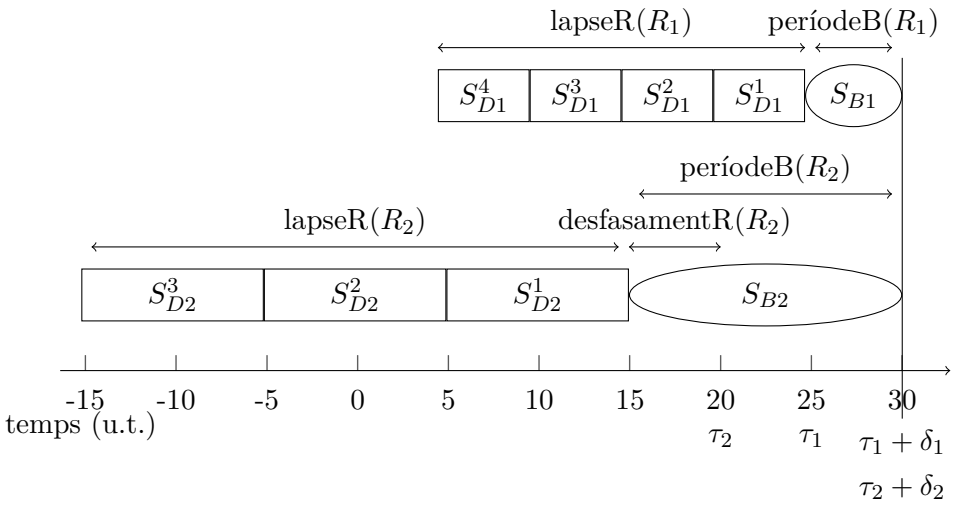


Figura 5.8.: Cronograma d'un esquema multiresolució just abans de la consolidació, on períodeB(R_i) = lapseB(R_i)

Si definim l'interval de temps del lapse com a $[\tau - \kappa\delta, \tau]$ i l'interval de temps real de la sèrie temporal del disc com a $[\min(S_D), \max(S_D)]$, aleshores normalment es complirà que $\max(S_D) = \tau$ i $\min(S_D) = \tau - (\kappa - 1)\delta$. No obstant això, pot no complir-se per exemple si la sèrie temporal no és regular o per exemple si τ i $\max(S_D)$ no coincideixen. Aquest darrer fet que τ i $\max(S_D)$ no coincideixen ocorre quan la funció d'agregació d'atributs causa un desfasament; ho anomenem desfasament de la subsèrie resolució.

Definició 5.22 (Desfasament de la subsèrie resolució)**.** *Sigui* $R = (S_B, S_D, \tau, \delta, \kappa, f)$ *una subsèrie resolució, el seu desfasament* desfasamentR(R) *és una durada de temps que indica la distància entre* τ *i* $\max(S_D)$ *causada per la funció d'agregació* f*. Així havent definit la consolidació sobre* $m = f(S, [\tau, \tau + \delta]$ *(v. definició 5.12), una funció d'agregació amb desfasament retorna una mesura amb* $T(m) = \tau -$ desfasamentR(R)*.*

Un altre desfasament que es produeix, aquest però variable, és la variació de temps que hi ha entre el darrer instant de consolidació i l'instant de temps actual. Aquest interval de temps és durant el qual el buffer emmagatzema les noves mesures que arriben i l'anomenen període de buffer.

Definició 5.23 (Període de buffer de la subsèrie resolució)**.** *Sigui* $R = (S_B, S_D, \tau, \delta, \kappa, f)$ *una subsèrie resolució,* desfasamentR(R) *el seu desfasament i* t *l'instant de temps actual, el període de buffer de la subsèrie resolució* perí̇odeB(R) *és una durada de temps que indica la distància entre* τ *i* t *tenint en compte el desfasament:* perí̇odeB(R) $= t - (\tau -$ desfasamentR(R))*.*

Si la consolidació de la subsèrie resolució es realitza immediatament cada cop que estigui en estat de consolidable i existeix una mesura $m \in S_B$ que compleix $T(m) = t$, aleshores el període de buffer té una variació afitada atès que, per ser la consolidació immediata, $t - \tau \leq \delta$. En aquest cas, el mínim que pot prendre el període de buffer és el desfasament, $\text{desfasamentR}(R) \leq \text{períodeB}(R)$, que correspon a l'instant $t = \tau$. El màxim que pot prendre l'anomenem lapse de buffer de la subsèrie resolució.

Definició 5.24 (Lapse de buffer de la subsèrie resolució)**.** *Sigui* $R = (S_B, S_D, \tau, \delta, \kappa, f)$ *una subsèrie resolució i* $\text{desfasamentR}(R)$ *el seu desfasament, el lapse de buffer de la subsèrie resolució* $\text{lapseB}(R)$ *és una durada de temps que indica el període de buffer màxim que pot prendre:* $\text{lapseB}(R) = \delta + \text{desfasamentR}(R)$*. Sempre es compleix que* $\text{períodeB}(R) \leq \text{lapseB}(R)$ *atès que, per ser la consolidació immediata,* $t - \tau \leq \delta$*.*

Una altra propietat de l'esquema de multiresolució és la regularització de les sèries temporals emmagatzemades en el disc. El període d'aquesta sèrie temporal del disc normalment es correspondrà amb δ. Així, segons el pas de consolidació descrit a l'esquema de multiresolució, descrivim quina subsèrie resolució conté més resolució.

Definició 5.25 (Subsèrie resolució amb més resolució)**.** *Siguin* $R_1 = (S_{B1}, S_{D1}, \tau_1, \delta_1, \kappa_1, f_1)$ *i* $R_2 = (S_{B2}, S_{D2}, \tau_2, \delta_2, \kappa_2, f_2)$ *dues subsèries resolució. La subsèrie amb més resolució,* $\text{maxR}(R_1, R_2)$*, és la que té el pas de consolidació més petit:* $\text{maxR}(R_1, R_2) = R_i$ *on* $\delta_i = \min(\delta_1, \delta_2)$*.*

Això no obstant, notem que en determinats instants el pas de consolidació descrit a l'esquema de multiresolució pot no coincidir amb el període de regularitat de la sèrie temporal emmagatzemada al disc, com per exemple durant un canvi en l'esquema del pas de consolidació (vegeu més endavant la definició 5.27).

Resumint, podem dir que una subsèrie resolució té, per una banda, informació consolidada per un lapse temporal de $\text{lapseR}(R)$, el qual es posiciona absolutament des de $\tau + \text{desfasamentR}(R)$ enrere. Per altra banda, la subsèrie resolució té informació no consolidada al davant de la consolidada per un interval de mida $\text{períodeB}(R)$ i de com a màxim $\text{lapseB}(R)$. A continuació mostrem aquestes propietats en diversos exemples mitjançant cronogrames d'esquemes multiresolució per a la mateixa sèrie temporal multiresolució però en diferents situacions temporals.

Exemple 5.4 (Instant de temps just abans de la consolidació)**.** El cronograma de la figura 5.8 mostra la sèrie temporal multiresolució $M = \{R_1, R_2\}$ de l'exemple 5.3 a l'instant de temps determinat, el $t = 30$. Exactament mostra una fotografia en l'instant just abans d'aplicar l'operació de consolidació per a totes dues subsèries, les quals ja són consolidables: $t \geq \delta_1 + \tau_1$ i $t \geq \delta_2 + \tau_2$ assumint que existeixen les mesures $m_1 \in S_{D1} : T(m_1) = t$ i $m_2 \in S_{D2} : T(m_2) = t$. Aquest és un moment en

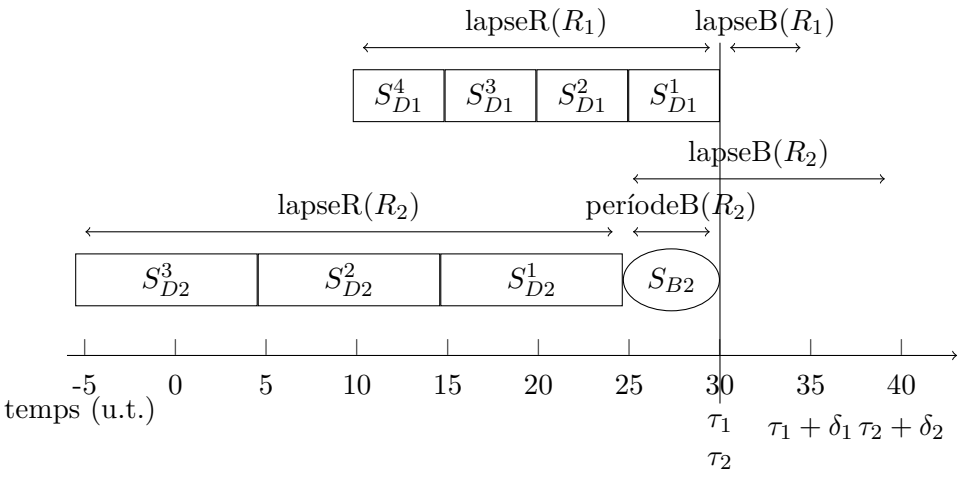


Figura 5.9.: Cronograma d'un esquema multiresolució just després de la consolidació, on períodeB(R_i) = desfasamentR(R_i)

què es compleix la propietat períodeB(R_i) = lapseB(R_i), és a dir en què el període de buffer té la mida màxima.

En aquest cronograma es poden observar totes les propietats d'un cronograma multiresolució que ara, un cop definides, en calculem els valors concrets per l'exemple:

- Lapses de les subsèries: lapseR(R_1) = $\kappa_1\delta_1 = 20$ i lapseR(R_2) = $\kappa_2\delta_2 = 30$.
- Desfasaments de les subsèries: desfasamentR(R_1) = 0 i desfasamentR(R_2) = 5, segons interpretació de la f_1 = mitjana i f_2 = mitjanad5.
- Períodes de buffer: períodeB(R_1) = $t-(\tau_1-$desfasamentR(R_1)) = $30-(25-0) = 5$ i períodeB(R_2) = $t-(\tau_2-$desfasamentR(R_2))) = $30-(20-5) = 15$ atès que en l'instant de temps actual $t = 30$ just abans de consolidar-se $\tau_1 = 25$ i $\tau_2 = 20$.
- Lapses de buffer: lapseB(R_1) = δ_1 + desfasamentR(R_1) = 5 i lapseB(R_2) = δ_2 + desfasamentR(R_2) = 15.
- Subsèrie amb més resolució: maxR(R_1, R_2) = R_1 atès que $\delta_1 = \min(\delta_1, \delta_2)$.

Exemple 5.5 (Instant de temps just després de la consolidació)**.** El cronograma canvia amb el pas del temps i per tant també canvia la mida dels buffers i la posició temporal dels discs. Ara el cronograma de la figura 5.9 mostra el mateix cas que l'exemple 5.4 per $t = 30$ però exactament en una fotografia a l'instant just després d'aplicar l'operació de consolidació per a totes dues subsèries. Aquest és un moment en què es compleix la propietat períodeB(R_i) = desfasamentR(R_i), és a dir en què el període de buffer té la mida mínima.

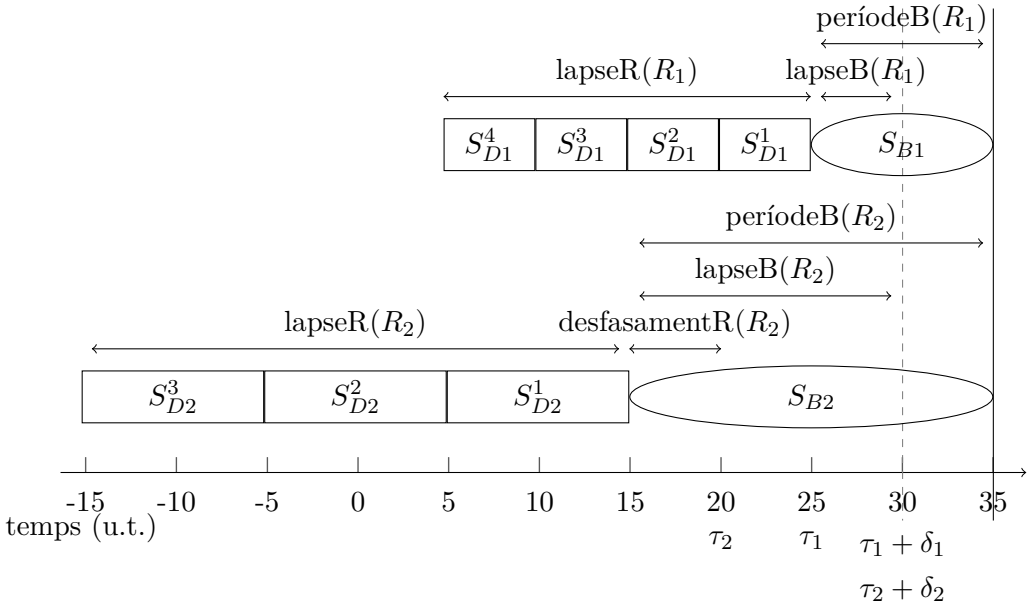


Figura 5.10.: Cronograma d'un esquema multiresolució amb consolidació retardada, on períodeB(R_i) > lapseB(R_i)

Pel que fa als valors de les propietats del cronograma, tots són els mateixos que a l'exemple 5.4 llevat de:

- Períodes de buffer: períodeB(R_1) = $t - (\tau_1 - \text{desfasamentR}(R_1)) = 30 - (30 - 0) = 0$ i períodeB(R_2) = $t - (\tau_2 - \text{desfasamentR}(R_2))) = 30 - (30 - 5) = 5$ atès que en l'instant de temps actual $t = 30$ just després de consolidar-se $\tau_1 = 30$ i $\tau_2 = 30$.

Exemple 5.6 (Cas no ideal amb consolidació retardada)**.** L'exemple 5.5 és la continuació temporal (instantània) de l'exemple 5.4 assumint que en l'instant 30 s'aplica l'operació de consolidació. Ara descrivim un cas no ideal en què aquesta operació de consolidació no s'aplica fins a l'instant $t = 35$. Això podria ser degut que no es vol executar l'operació en el mateix moment que les subsèries siguin consolidables o bé que encara no són consolidables perquè no existeixen les mesures $m_1 \in S_{B1} : 30 \leq T(m_1) \leq 35$ i $m_2 \in S_{B2} : 30 \leq T(m_2) \leq 35$.

Així doncs, reprenem l'exemple 5.4 i avancem el temps actual fins a l'instant $t = 35$ però sense aplicar la consolidació. El cronograma es mostra a la figura 5.10, en línia contínua vertical es mostra l'instant de temps actual 30 i en línia discontínua es mostra la consolidació immediata ideal que hauria estat a l'instant 30. Aquest és un moment en què apareix la propietat períodeB(R_i) > lapseB(R_i), és a dir en què el període de buffer supera la mida màxima que té en cas de consolidació immediata.

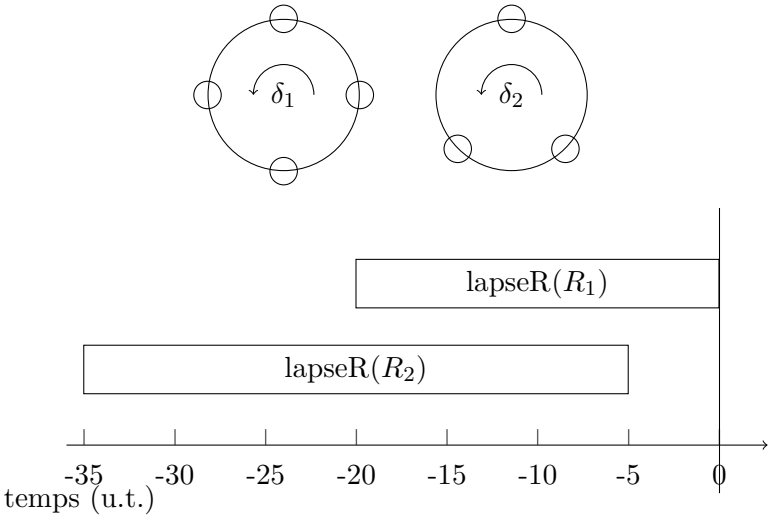


Figura 5.11.: Cronograma periòdic d'un esquema multiresolució

Pel que fa als valors de les propietats del cronograma, tots són els mateixos que a l'exemple 5.4 llevat de:

- Períodes de buffer: períodeB(R_1) $= t - (\tau_1 - \text{desfasamentR}(R_1)) = 35 - (25 - 0) = 10$ i períodeB(R_2) $= t - (\tau_2 - \text{desfasamentR}(R_2))) = 35 - (20 - 5) = 20$ atès que en l'instant de temps actual $t = 35$ encara queda pendent de consolidar-se $\tau_1 = 25$ i $\tau_2 = 20$.

Cronograma periòdic El cronograma evoluciona amb el temps i a cada instant determinat presenta una forma. Això no obstant totes les propietats de l'esquema són fixes llevat del període de buffer que varia segons t, a més si assumim el cas ideal de consolidació immediata aleshores les formes es repeteixen periòdicament en el temps de δ per a cada subsèrie resolució. Per a simplificar, dibuixem un cronograma on no aparegui el període de buffer i que només mostri la informació periòdica del cronograma. A tal efecte, establim l'instant 0 com aquell on totes les subsèries resolució es consoliden al mateix temps i dibuixem els lapses enrere en el temps. A la figura 5.11 es pot veure el cronograma periòdic per als exemples anteriors on l'eix del temps indica les durades i les posicions relatives dels discs de les subsèries resolució. La informació referent al pas de consolidació i la resolució dels discs es dibuixa a la part superior amb formes circulars, de manera que s'observa clarament quina subsèrie té més resolució.

Canvis en l'esquema

Canviar l'esquema multiresolució d'una sèrie temporal significa crear un esquema nou de multiresolució i emmagatzemar-hi les dades de l'esquema vell de forma que es conservi la coherència que tenien aquestes dades. Així doncs, a continuació definim algunes operacions de canvi d'esquema que es poden aplicar als objectes d'una base de dades multiresolució de forma coherent amb les dades emmagatzemades.

L'operació que canvia la mida del disc d'una subsèrie resolució ha de controlar que si la mida disminueix s'han d'eliminar dades.

Definició 5.26 (Canvi de mida d'una subsèrie resolució)**.** *Sigui* $R = (B, D)$ *una subsèrie resolució, en què* $D = (S_D, \kappa)$ *és el disc de la subsèrie, i* κ' *un nou cardinal màxim, el canvi de mida de la subsèrie resolució* canviaK(R, κ') *és una nova subsèrie resolució* $R' = (B, D')$ *amb les dades antigues afegides de nou al disc:* canviaK$(R, \kappa') = (B, D')$ *aplicant l'operació* $\forall m_i \in S_D$: afegeixD(D', m_i) *amb* $D' = (\{\}, \kappa')$ *com a disc inicial.*

L'operació que canvia el pas de consolidació d'una subsèrie resolució no cal que tingui en compte les dades ja que la sèrie temporal emmagatzemada s'anirà canviant quan es consolidin noves mesures.

Definició 5.27 (Canvi de pas de consolidació d'una subsèrie resolució)**.** *Sigui* $R = (S_B, S_D, \delta, \tau, \kappa, f)$ *una subsèrie resolució i* δ' *un nou pas de consolidació, el canvi del pas de consolidació de la subsèrie resolució* canvia$\delta(R, \delta')$ *és la nova subsèrie resolució* canvia$\delta(R, \delta') = (S_B, S_D, \delta', \tau, \kappa, f)$.

Hi ha altres canvis de l'esquema que tampoc no requereixen tenir en compte les dades, com per exemple canviar la funció d'agregació d'atribut. Atès que són molt semblants al canvi de pas de consolidació, no mostrem les definicions específics per a aquestes operacions.

L'operació que canvia la resolució d'una subsèrie resolució modifica alhora la mida i el pas de consolidació tenint en compte les dades; és a dir, canvia el període de la sèrie temporal emmagatzemada seguint el criteri de representació.

Definició 5.28 (Canvi de resolució d'una subsèrie resolució)**.** *Sigui* $R = (S_B, S_D, \delta, \tau, \kappa, f)$ *una subsèrie resolució,* r *una funció de representació per a la sèrie temporal,* κ' *un nou cardinal màxim i* δ' *un nou pas de consolidació, el canvi de resolució de la subsèrie resolució* canviaResolució(R, κ', δ', r) *és una nova subsèrie resolució* $R' = (S_B, S'_D, \delta', \tau, \kappa', f)$ *amb una selecció temporal segons el criteri de representació en el nou conjunt regular de temps:* canviaResolució$(R, \kappa', \delta', r) = (S_B, S'_D, \delta', \tau, \kappa', f)$ *a on* $S'_D = S_D[T]^r$ *i* $T = \{\tau - n\delta' | n \in \mathbb{N}, n < \kappa'\}$.

Per a treballar amb sèries temporals multivaluades cal que les sèries temporals emmagatzemades als buffers i al discs tinguin la mateixa forma. Així afegir un nou multivalor significa afegir un nou atribut a les sèries temporals del buffer i dels disc de cada subsèrie resolució. A continuació definim com afegir un multivalor a una subsèrie resolució considerant sèries temporals en la forma canònica; a la pràctica, per comoditat, habitualment cada atribut de multivalor tindrà un nom.

Definició 5.29 (Afegeix multivalor a una subsèrie resolució)**.** *Sigui* $R = (S_B, S_D, \delta, \tau, \kappa, f)$ *una subsèrie resolució, l'addició d'un nou multivalor és una subsèrie resolució* afegeixMultivalor$(R) = (S'_B, S'_D, \delta, \tau, \kappa, f)$ *amb les sèries temporals ampliades amb un nou atribut inicialment de valor indefinit* $S'_B = \text{mapa}(S_B, g)$ *on* $g(t, v) = (t, (v, \infty))$ *i* $S'_D = \text{mapa}(S_D, h)$ *on* $h(t, v) = (t, (v, \infty))$.

Per a eliminar un multivalor només cal eliminar l'atribut a totes les sèries temporals del buffer i dels disc de cada subsèrie resolució. Atès que és més senzill que afegir un multivalor, no en mostrem la definició específica.

Unió i junció de dos esquemes

Dos casos particulars de canvi d'esquema multiresolució són el d'unió i el de junció de dos esquemes. En aquests canvis cal crear un nou esquema tot conservant la coherència de dades emmagatzemades en dos esquemes diferents.

Un exemple d'aplicació de la unió de dos esquemes és el següent. Es mesuren els valors d'una sèrie temporal i durant un temps s'emmagatzemen com a sèrie temporal multiresolució en una base de dades però durant un altre temps s'emmagatzemen en una altra base de dades. En acabar es volen unir els valors emmagatzemats a les dues bases de dades.

L'operació d'unió de dues subsèries resolució és una unió de les sèries temporals de cada un.

Definició 5.30 (Unió de dues subsèries resolució)**.** *Sigui* $R_1 = (S_{B1}, S_{D1}, \delta_1, \tau_1, \kappa_1, f_1)$ *i* $R_2 = (S_{B2}, S_{D2}, \delta_2, \tau_2, \kappa_2, f_2)$ *dues subsèries resolució, la unió de les dues subsèries resolució* unióR(R_1, R_2) *és una subsèrie resolució* $R' = (S'_B, S'_D, \delta', \tau', \kappa', f')$ *que conté la unió de les sèries temporals dels seus buffers i discs:* unióR$(R_1, R_2) = (S'_B, S'_D, \delta_1, \max(\tau_1, \tau_2), \kappa_1 + \kappa_2, f_1)$ *a on* $S'_B = S_{B1} \cup S_{B2}$ *i* $S'_D = S_{D1} \cup S_{D2}$.

La unió de dues subsèries resolució no és commutativa a causa que les unions de les sèries temporals no ho són. Tampoc ho és a causa que si s'uneixen dues subsèries resolució amb diferent δ i f llavors es determina que la primera marca quins són els δ' i f' resultants. És a dir que en cas que es vulguin unir subsèries resolució que continguin mesures en el mateix temps o informació diferent; es prioritza la primera.

L'operació d'unió de dues sèries temporals multiresolució és la unió de conjunts per a les subsèries resolució quan no intersequen en les claus (δ, f). En cas que intersequin cal unir les dues subsèries resolució.

Definició 5.31 (Unió de dues sèries temporals multiresolució)**.** *Sigui M_1 i M_2 dues sèries temporals multiresolució, les subsèries resolució de les quals tenen la forma $R_i = (S_{Bi}, S_{Di}, \delta_i, \tau_i, \kappa_i, f_i)$. La unió de les dues* unióM(M_1, M_2) *retorna una sèrie temporal multiresolució que conté les subsèries que no intersequen i la unió de les subsèries que intersequen:*

$$\text{unióM}(M_1, M_2) = M \cup M_1' \cup M_2'$$

on

$$M = \{\text{unióR}(R_1, R_2) | R_1 \in M_1 \wedge R_2 \in M_2 \wedge (\delta_1, f_1) = (\delta_2, f_2)\}$$
$$M_1' = \{R_1 | R_1 \in M_1 \wedge \nexists R_2 \in M_2 : (\delta_1, f_1) = (\delta_2, f_2)\}$$
$$M_2' = \{R_2 | R_2 \in M_2 \wedge \nexists R_1 \in M_1 : (\delta_1, f_1) = (\delta_2, f_2)\}$$

La unió de dues sèries temporals multiresolució no és commutativa a causa que la unió de subsèries resolució no ho és.

L'operació de junció de dues subsèries resolució és la fusió de les dades de les dues en una sèrie temporal multivalor. En el cas que els vectors de temps siguin els mateixos es pot utilitzar la junció de sèries temporals. En el cas que alguns temps no coincideixin cal utilitzar la junció temporal amb una representació que donarà valors allà on manquin.

Definició 5.32 (Junció de dues subsèries resolució)**.** *Sigui $R_1 = (S_{B1}, S_{D1}, \delta_1, \tau_1, \kappa_1, f_1)$ i $R_2 = (S_{B2}, S_{D2}, \delta_2, \tau_2, \kappa_2, f_2)$ dues subsèries resolució i r una funció de representació de sèries temporals, la junció de les dues subsèries resolució* junció$\text{R}^r(R_1, R_2)$ *és una subsèrie resolució $R' = (S_B', S_D', \delta', \tau', \kappa', f')$ que conté la junció temporal de les sèries temporals dels seus buffers i discs:* junció$\text{R}^r(R_1, R_2) = (S_B', S_D', \delta_1, \max(\tau_1, \tau_2), \kappa_1 + \kappa_2, f_1)$ *a on $S_B' = S_{B1} \bowtie^r S_{B2}$ i $S_D' = S_{D1} \bowtie^r S_{D2}$.*

La junció de dues subsèries resolució no és commutativa a causa que si s'uneixen dues subsèries resolució amb diferent δ i f llavors es determina que la primera marca quins són els δ' i f' resultants.

L'operació de junció de dues sèries temporals multiresolució és la unió de conjunts per a les subsèries resolució, ampliades amb un multivalor indefinit, quan no intersequen en les claus (δ, f). En cas que intersequin cal ajuntar les dues subsèries resolució.

Definició 5.33 (Junció de dues sèries temporals multiresolució)**.** *Sigui M_1 i M_2 dues sèries temporals multiresolució, les subsèries resolució de les quals tenen la forma $R_i = (S_{Bi}, S_{Di}, \delta_i, \tau_i, \kappa_i, f_i)$, i sigui r una funció de representació de sèries*

temporals. La junció de totes dues, junció$\mathrm{M}^r(M_1, M_2)$*, retorna una sèrie temporal multiresolució que conté les subsèries que no intersequen ampliades amb un multivalor i la junció de les subsèries que intersequen:*

$$\text{junció}\mathrm{M}^r(M_1, M_2) = M \cup M_1' \cup M_2'$$

on

$$M = \{\text{junció}\mathrm{R}(R_1, R_2) | R_1 \in M_1 \wedge R_2 \in M_2 \wedge (\delta_1, f_1) = (\delta_2, f_2)\}$$
$$M_1' = \{\text{afegeixMultivalor}(R_1) | R_1 \in M_1 \wedge \nexists R_2 \in M_2 : (\delta_1, f_1) = (\delta_2, f_2)\}$$
$$M_2' = \{\text{afegeixMultivalor}(R_2) | R_2 \in M_2 \wedge \nexists R_1 \in M_1 : (\delta_1, f_1) = (\delta_2, f_2)\}$$

5.2.3. Consultes

En els SGBD les consultes són les operacions que permeten treballar amb les dades emmagatzemades. En el cas dels SGSTM les dades són sèries temporals emmagatzemades amb forma de multiresolució. Així doncs, els operadors de consulta que es defineixen a continuació permeten extreure les sèries temporals que hi ha emmagatzemades amb l'objectiu d'aplicar-hi posteriorment els operadors propis de les sèries temporals, com s'ha definit en el model dels SGST.

Els operadors de consulta en un SGSTM es basen en obtenir sèries temporals de la base de dades multiresolució per a aplicar-hi operacions dels SGST. En aquest apartat definim el operadors que permeten obtenir sèries temporals a partir d'una sèrie temporal multiresolució; l'aplicació posterior d'operadors a les sèries temporals segueix les mateixes definicions que en el model d'operacions dels SGST (v. secció 4.2).

Distingim entre dos tipus d'operadors de consulta. Uns permeten extreure les subsèries temporals de les subsèries resolució i uns altres permeten abstreure una sèrie temporal resolució com si fos una sola sèrie temporal amb diferents períodes de mostreig.

Extracció de subsèries

Les subsèries resolució estan caracteritzades per la parella de pas de consolidació i funció d'agregació d'atributs (δ, f), els quals són els atributs clau del conjunt. Seleccionant aquests dos paràmetres d'una sèrie temporal multiresolució s'obté la subsèrie resolució corresponent, de la qual es pot extreure la sèrie temporal emmagatzemada al seu buffer o al seu disc. A continuació mostrem con seleccionar la sèrie temporal del disc, per a la del buffer es pot procedir de forma similar.

Definició 5.34 (Selecció de la sèrie temporal d'un disc)**.** *Sigui M una sèrie temporal multiresolució i sigui (δ, f) una parella d'atributs clau. La selecció de la sèrie temporal d'un disc de la sèrie temporal multiresolució és una sèrie temporal tal que $\exists(B, D) \in M : B = (S, \tau, \delta, f) \wedge D = (\text{SèrieDisc}(M, \delta, f), \kappa)$ on S, τ i κ són variables lligades. Noteu que hem assumit que no hi ha parelles (δ, f) repetides en una sèrie temporal multiresolució.*

Sèrie temporal total

La sèrie temporal total és la sèrie temporal que ofereix la màxima resolució de la sèrie temporal multiresolució; és a dir que concatena les subsèries resolució tenint en compte el pas de consolidació de cada una. El resultat és una sèrie temporal amb períodes de mostreig regulars a trossos.

Definició 5.35 (Sèrie temporal total)**.** *Sigui $M = \{R_0, \ldots, R_k\}$ una sèrie temporal multiresolució i siguin $S_0, \ldots, S_k$ les sèries temporals de cadascun dels discs i δ_0, $\ldots, \delta_k$ els corresponents passos de consolidació. Assumiu que M està ordenada en una permutació que compleix $\delta_0 < \cdots < \delta_k$. La sèrie temporal total de la sèrie temporal multiresolució,* $\text{SèrieTotal}(M)$*, retorna la sèrie temporal que resulta de la concatenació de les sèries temporals dels discs per ordre de pas de consolidació:* $\text{SèrieTotal}(M) = S_0 || \cdots || S_k$.

Per tal que els passos de consolidació de la sèrie temporal multiresolució tinguin un ordre estricte, $\delta_0 < \cdots < \delta_k$, cal que no hi hagi δ repetits. En el cas que una sèrie temporal multiresolució tingui δ repetits, prèviament a l'obtenció de la sèrie temporal total cal decidir com seleccionar un conjunt únic de δ. Proposem dos exemples de selecció prèvia:

- Una sèrie temporal multiresolució és un conjunt amb (δ, f) com a atributs clau i una selecció prèvia habitual pot ser la selecció de les subsèries resolució que comparteixin un determinat agregador d'atributs f.
- Un altra selecció prèvia possible és utilitzar l'operació d'unió de subsèries resolució per a les subsèries que tinguin el mateix pas de consolidació δ.

L'operació de sèrie temporal total definida és una operació genèrica per a extreure una sèrie temporal amb la màxima resolució, però se'n poden definir altres casos particulars. Per exemple, una operació de sèrie temporal total on la concatenació sigui temporal: $S_0 ||^r \cdots ||^r S_k$; o una on es consulti una subsèrie resolució en particular $S_i = \text{SèrieDisc}(M, \delta, f)$ i s'acabi de completar la informació amb les dades d'altres resolucions: $S_i ||^r S_0 ||^r \cdots ||^r S_k$.

Així doncs, la sèrie temporal total és una abstracció d'una sèrie temporal multiresolució en forma de sèrie temporal. Aleshores es poden aplicar totes les operacions de les sèries temporals dels SGST. En mostrem dos exemples:

- Per a extreure una resolució determinada de la sèrie temporal multiresolució M, es consulta la sèrie temporal total i s'aplica una selecció de resolució: SèrieTotal$(M)[T]^r$ on T és un conjunt d'instants de temps i r la representació de la sèrie temporal.
- Per a sumar dues sèries temporals emmagatzemades com a sèries temporals multiresolució, es consulta la sèrie temporal total de cada una i s'hi aplica l'operació computacional de sumar (v. ex. 4.17): SèrieTotal(M_1) + SèrieTotal(M_2).

5.3. Funcions d'agregació d'atributs

Les funcions d'agregació d'atributs s'utilitzen en la consolidació dels buffers per tal de compactar certa informació de la sèrie temporals. Sigui S una sèrie temporal, τ un instant de consolidació i δ un pas de consolidació. Una funció d'agregació d'atributs f calcula una nova mesura $m = f(S, \tau, \delta)$. A partir de τ i δ s'obté l'interval de temps $[\tau, \tau + \delta]$. Aleshores, la mesura resultant m s'interpreta com un resum de la informació de S per l'interval de temps $[\tau, \tau + \delta]$.

Generalment, m resulta d'aplicar dues operacions a S:

1. una selecció d'una subsèrie S' segons l'interval de temps $[\tau, \tau+\delta]$, per exemple $S' = S[\tau, \tau + \delta]$,
2. i una agregació amb aquesta subsèrie $m = \text{agregació}(S', n, g)$ on n és una mesura i g una funció (v. def. 4.32) que computen l'agregació seguint el criteri de f.

Atès que hi ha maneres diferents de resumir la informació d'una sèrie temporal, cal plantejar diferents funcions d'agregació d'atributs. Per exemple, es poden calcular estadístics de la sèrie temporal, com el valor màxim o la mitjana; aplicar operacions de processament digital del senyal, com fan Zhang et al. [139], o algoritmes per a detectar comportaments aberrants, com fa Brutlag [11]. A més a més, la representació de les sèries temporals (v. apartat 4.3.2) pot afectar els càlculs que es fan en l'agregació o bé es pot aprofitar l'agregació per a tractar algunes de les patologies de les sèries temporals (v. apartat 4.3.3). Així doncs, es poden definir una enorme varietat de funcions d'agregació d'atributs i no hi ha cap assumpció global que es pugui fer, cada usuari ha d'interpretar quina combinació d'agregació i representació s'adiu més amb el fenomen mesurat. Com a conseqüència, els SGSTM han de donar llibertat als usuaris per a definir funcions d'agregació d'atributs personalitzades.

Com a mostra de com dissenyar funcions d'agregació d'atributs, a continuació descrivim algunes interpretacions possibles que se'n poden fer, tant pel que fa al càlcul de l'instant de temps resultant de la consolidació com pel que fa al càlcul amb representació de sèries temporals, i descrivim com utilitzar-les per a tractar i validar dades desconegudes en les sèries temporals.

5.3.1. Interpretació de l'agregació

L'agregació d'una sèrie temporal en un interval resulta en una mesura m. Així per a definir les operacions d'agregació cal interpretar quin ha de ser el temps resultant $T(m)$ i el valor resultant $V(m)$.

Podem definir patrons generals de funcions d'agregació d'atributs que indiquin quina informació o estadístic es resumeix de la sèrie temporal, és a dir patrons generals que indiquin com s'ha de calcular el valor resultant $V(m)$ independentment del mètode de representació que es vulgui associar a la sèrie temporal. Tot i així, el temps resultant $T(m)$ no queda definit sinó que s'ha interpretar coherentment per a cada cas particular de representació.

Sigui f una funció d'agregació d'atributs, sigui $m = f(S, \tau, \delta)$ el càlcul de la mesura resultant i sigui $S^r(t)$ una funció de representació de la sèrie temporal S en què t es calcularà a partir de τ i δ. A continuació mostrem alguns exemples de patrons generals per a calcular el valor resultant $V(m)$:

- El *màxim* calcula $V(m)$ com $V(m) = \max_{\forall t \in [\tau, \tau+\delta]} S^r(t)$. Resumeix S amb el màxim dels valors de les mesures a l'interval $[\tau, \tau + \delta]$.
- El *darrer* calcula $V(m)$ com $V(m) = S^r(\tau + \delta)$. Resumeix S amb el valor del darrer instant de temps de l'interval $[\tau, \tau + \delta]$.
- La *mitjana* calcula $V(m)$ com $V(m) = \frac{1}{\delta} \int_{\tau}^{\tau+\delta} S^r(t)dt$. Resumeix S amb la *mitjana de la funció* [132] a l'interval $[\tau, \tau + \delta]$.

En aquests patrons d'atributs es treballa sobre una funció $S^r(t)$, que a cada cas serà una funció de representació concreta. Això permetrà, per una banda, interpretar coherentment el temps resultant $T(m)$. Per altra banda, permetrà interpretar amb matemàtica discreta el càlcul del valor resultant $V(m)$; noteu que els patrons anteriors s'han definit com a problemes d'anàlisi numèric però per a cada funció de representació concreta $S^r(t)$ podrem expressar els operadors segons el model de SGST descrit amb àlgebra discreta matemàtica. Així doncs, a continuació exemplifiquen algunes interpretacions possibles per al càlcul de $T(m)$ i de $V(m)$.

Temps d'agregació resultant

L'objectiu de les funcions d'agregació d'atributs és determinar un instant de temps $T(m)$ i un valor $V(m)$. Aquest càlcul del temps i del valor es pot realitzar al mateix temps però també pot ser independent. Així, en principi el temps resultant serà independent i valdrà $T(m) = \tau + \delta$ per estar d'acord amb l'operació de consolidació del buffer i no causar desfasament de la subsèrie resolució (v. definició 5.22). Però en alguns casos aquest $T(m)$ serà dependent del valor calculat o estarà subjecte a una interpretació adient com és el cas de les representacions a l'apartat següent.

Un exemple de funció d'agregació on temps i valor són dependents és una funció que retorni la primera mesura que troba, primera$(S, \tau, \delta) = \min(S[\tau, \tau + \delta))$. Llavors el temps resultant pot ser $\tau \leq T(m) < \tau + \delta$. Noteu, però, que en aquest cas la sèrie temporal consolidada resultant no és regular perquè els temps resultants depenen de les mesures de cada interval.

Un exemple de funció d'agregació on temps i valor són independents i on la subsèrie resolució resultant és regular però amb desfasament, és una funció que fa la mitjana amb un desfasament de 5 unitats de temps. La funció d'agregació mitjanad5 s'ha utilitzat anteriorment a l'exemple 5.3, ara podem definir-la contextualitzada en les funcions d'agregació d'atributs: mitjanad5$(S, \tau, \delta) = m$ on $V(m) =$ mitjana_v$(S[\tau - 5, \tau + \delta - 5))$ i $T(m) = \tau + \delta - 5$.

Agregació amb representació

La varietat de funcions de representació per les sèries temporals indueix a una varietat de funcions d'agregació per a un mateix patró d'atributs. Per exemple, la funció d'agregació per l'atribut de màxim dóna com a resultat valors diferents si es considera una representació lineal o una representació a trossos constant. A continuació mostrem la interpretació dels patrons definits anteriorment per a tres mètodes de representació: parcial discreta (PD), delta de Dirac (DD) i zero-order hold enrere (ZOHE).

Parcial discreta. En els casos parcials, $S^r(t)$ no és totalment contínua en el temps, però es pot resoldre l'agregació del valor resultant assumint que el domini de temps $\mathcal{T}$ es correspon als instants de temps que hi ha a la sèrie temporal, és a dir $\mathcal{T} = \Pi_t(S)$. El temps resultant es pot interpretar segons s'ha descrit a l'apartat anterior, per exemple $T(m) = \tau + \delta$. A més, també es pot interpretar l'interval de temps d'agregació $[\tau, \tau + \delta]$. Així sigui S la sèrie original, el resultat es pot calcular sobre una subsèrie amb interval obert $S' = S(\tau, \tau + \delta)$, tancat $S' = S(\tau, \tau + \delta]$, semiobert $S' = S(\tau, \tau + \delta]$ o $S' = S[\tau, \tau + \delta)$, o altres combinacions com per exemple tenir desfasaments $S' = S[\tau - d, \tau + \delta - d]$ on d és una durada. Així de forma general podem definir les funcions d'agregació d'atributs amb representació PD com $f(S, \tau, \delta) = m$ on $T(m) = \tau + \delta$ i el valor resultant $V(m)$ depèn del l'atribut que es vulgui resumir calculat en l'interval $S' = S[\tau, \tau + \delta]$, a continuació es mostren els patrons d'exemple interpretats segons aquest criteri.

Definició 5.36 (Agregació parcial discreta)**.** *Sigui S una sèrie temporal, τ un instant de consolidació, δ un pas de consolidació i $S' = S[\tau, \tau + \delta]$ un interval de la sèrie temporal. Les funcions d'agregació* PD *per als atributs màxim, darrer i mitjana són:*

- màxim$^{\text{PD}}(S, \tau, \delta) = m$ *on* $V(m) = \max_{\forall m' \in S'}(V(m'))$ *i* $T(m) = \tau + \delta$. *Aquest càlcul de* $V(m)$ *es correspon amb l'operació* max_v(S') *dels* SGST.
- darrer$^{\text{PD}}(S, \tau, \delta) = m$ *on* $V(m) = V(\max(S'))$ *i* $T(m) = \tau + \delta$.
- mitjana$^{\text{PD}}(S, \tau, \delta) = m$ *on* $V(m) = \frac{1}{|S'|} \sum_{\forall m' \in S'} V(m')$ *i* $T(m) = \tau + \delta$. *Aquest càlcul de* $V(m)$ *es correspon amb l'operació* mitjana_v(S') *dels* SGST*, és a dir amb calcular la mitjana aritmètica dels valors de les mesures.*

Delta de Dirac. Per a les funcions d'agregació delta de Dirac interpretem el temps d'agregació resultant centrat en l'interval $T(m) = \frac{2\tau+\delta}{2}$, tot i que també es podrien considerar altres interpretacions com per exemple $T(m) = \tau + \delta$. Així de forma general podem definir les funcions d'agregació d'atributs amb representació DD com $f(S, \tau, \delta) = m$ on $T(m) = \frac{2\tau+\delta}{2}$ i el valor resultant $V(m)$ depèn del l'atribut que es vulgui resumir calculat en l'interval temporal DD $S' = S[\tau, \tau + \delta]^{\text{DD}}$.

Definició 5.37 (Agregació delta de Dirac)**.** *Sigui S una sèrie temporal, τ un instant de consolidació, δ un pas de consolidació i $S' = S[\tau, \tau + \delta]^{\text{DD}}$ un interval temporal de la sèrie temporal. Les funcions d'agregació* DD *per als atributs màxim, darrer i mitjana són:*

- $\text{màxim}^{\text{DD}}(S, \tau, \delta) = m$ *on* $V(m) = \max\left(0, \max_{\forall m' \in S'}(V(m'))\right)$ *i* $T(m) = \frac{2\tau+\delta}{2}$.
- $\text{darrer}^{\text{DD}}(S, \tau, \delta) = m$ *on* $V(m) = V(\max(S'))$ *i* $T(m) = \frac{2\tau+\delta}{2}$.
- $\text{mitjana}^{\text{DD}}(S, \tau, \delta) = m$ *on* $V(m) = \frac{1}{\delta} \sum_{\forall m' \in S'} V(m')$ *i* $T(m) = \frac{2\tau+\delta}{2}$. *Nota: la funció delta de Dirac té la propietat fonamental* $\int \delta(t)dt = 1$.

Zero-order hold enrere. Per a les funcions d'agregació ZOHE interpretem sempre el temps d'agregació resultant com $T(m) = \tau + \delta$, atès que la representació ZOHE es defineix amb funcions graó contínues per l'esquerra. Així de forma general podem definir les funcions d'agregació d'atributs amb representació ZOHE com $f(S, \tau, \delta) = m$ on $T(m) = \tau + \delta$ i el valor resultant $V(m)$ depèn de l'atribut que es vulgui resumir calculat en l'interval temporal ZOHE $S' = S[\tau, \tau + \delta]^{\text{ZOHE}}$.

Definició 5.38 (Agregació zero-order hold enrere)**.** *Sigui S una sèrie temporal, τ un instant de consolidació, δ un pas de consolidació i $S' = S[\tau, \tau + \delta]^{\text{ZOHE}}$ un interval temporal de la sèrie temporal. Les funcions d'agregació* ZOHE *per als atributs màxim, darrer i mitjana són:*

- $\text{màxim}^{\text{ZOHE}}(S, \tau, \delta) = m$ *on* $V(m) = \max_{\forall m' \in S'}(V(m'))$ *i* $T(m) = \tau + \delta$.
- $\text{darrer}^{\text{ZOHE}}(S, \tau, \delta) = m$ *on* $V(m) = V(\max(S'))$ *i* $T(m) = \tau + \delta$.
- $\text{mitjana}^{\text{ZOHE}}(S, \tau, \delta) = m$ *on* $V(m) = \frac{1}{\delta}\left[(T(o) - \tau)V(o) + \sum_{\forall m' \in S''} (T(m') - T(\text{ant}_S(m')))V(m')\right]$; $o = \min(S')$; $S'' = S' - \{o\}$; *i* $T(m) = \tau + \delta$.

Un cop definits els tres exemples de famílies d'agregacions, podem comparar-les en funció de com resumeixen la informació de la sèrie temporal. Reprenent la consolidació d'un buffer B (v. apartat 5.1.1), l'interval de consolidació es correspon a $[\tau, \tau + \delta]$ i és consolidable quan existeix una mesura $n \in B$ tal que $T(n) \geq \tau + \delta$. A la figura 5.12 dibuixem les mesures d'una sèrie temporal en vermell, un interval de consolidació del buffer en línies blaves i la mesura resultant de consolidació en

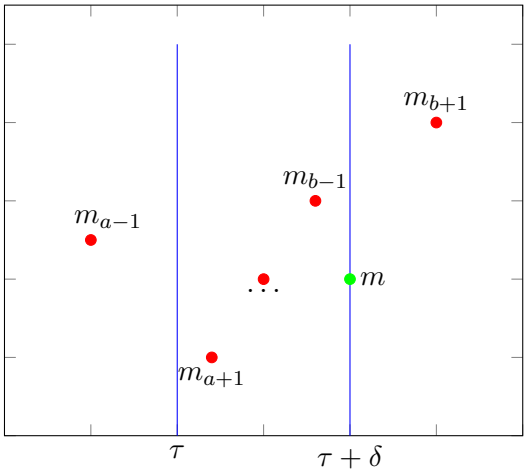


Figura 5.12.: Agregació d'un interval de la sèrie temporal

verd. Així, sigui $S = \{\ldots, m_{a-1}, m_{a+1}, \ldots, m_{b-1}, m_{b+1}, \ldots\}$ una sèrie temporal on $T(m_{a-1}) < \tau < T(m_{a+1}) < \ldots < T(m_{b-1}) < \tau + \delta < T(m_{b+1})$ i la consolidació del buffer que calcula la mesura resultant $m = f(S, \tau, \delta)$ amb la funció d'agregació d'atributs f. Assumim $T(m) = \tau + \delta$ per simplificar el dibuix, de manera general el càlcul del valor resultant és una agregació a partir de les mesures:

- $\{m_{a+1}, \ldots, m_{b-1}\}$ en el cas de les agregacions PD
- $\{(\tau, 0), (\ldots, 0), m_{a+1}, \ldots, (\ldots, 0), \ldots, m_{b-1}, (\ldots, 0), (t_b, 0)\}$ en el cas de les agregacions DD
- $\{m_{a+1}, \ldots, m_{b-1}, m_{b+1}\}$ en el cas de les agregacions ZOHE

En resum, alguns exemples mostrats de patrons tenen una interpretació semblant per a les representacions particulars, en certa manera només es diferencien en la interpretació de l'interval on s'ha de resumir la sèrie temporal. Per exemple la diferència principal en els atributs de màxim i darrer per a les tres representacions rau en la S', tot i que en el cas del màxim$^{\text{DD}}$ l'agregació a més ha de tenir en compte que en la funció de representació hi ha valors intermitjos que valen zero.

Altres exemples són molt diferents, com és el cas de l'atribut mitjana. En aquest cas, per a la PD i la DD és el càlcul de la suma dels valors tot i que dividit per $|S'|$ en la primera i per δ en la segona, i és una mitjana ponderada per les durades de temps en la ZOHE. En general, es pot dissenyar qualsevol operació d'agregació, com per exemple calcular la mitjana aritmètica de l'interval ZOHE amb mitjana_v($S[\tau, \tau + \delta]^{\text{ZOHE}}$), tot i que llavors cal interpretar quin patró d'atribut li correspon o altrament aquesta operació d'agregació pot no tenir sentit real.

Oetiker [97] utilitza a RRDtool una funció d'agregació semblant a la mitjana$^{\text{ZOHE}}$ per a resumir la informació conservant el comptatge total si les sèries temporals

mesurades tenen trets semàntics de comptador i són en forma de velocitat; així aquesta agregació es pot veure com una consolidació que conserva l'àrea del senyal original.

5.3.2. Tractament i validació de dades

En les patologies de les sèries temporals (v. apartat 4.3.3) s'ha descrit el problema de les dades desconegudes, les funcions d'agregació d'atributs poden cooperar en els processos de validació i tractament de dades. Així, les funcions d'agregació poden marcar o tractar dades desconegudes:

- Marcar dades com a desconegudes. És a dir determinar quan el resultat d'una agregació ha de ser desconegut perquè la sèrie temporal avaluada pateix una de les causes descrites: valors fora de rang, temps de termini excedit, etc.
- Tractar dades que són desconegudes, ja sigui perquè d'origen són desconegudes o perquè les hem marcat abans com a desconegudes. Si una funció d'agregació rep valors que són desconeguts, des d'un punt de vista estricte el resultat de l'agregació ha de ser desconegut. No obstant això, es poden aplicar operacions que tractin aquest valors desconeguts: reconstrucció del senyal, ignorar els valors desconeguts, etc.

A continuació definim el procés que fan les funcions d'agregació per a ambdós casos. Com a exemple de domini pels valors utilitzem els nombres reals projectius $\mathbb{R}^*$, en els quals representem el valor desconegut mitjançant l'element infinit (∞), segons la definició 4.5 de mesura de valor indefinit. Això no obstant, el domini de valors podria tenir diversos valors per a marcar diferents casos de dades desconegudes.

Tractament de dades desconegudes. Una funció d'agregació d'atributs f que tracti dades desconegudes és aquella que pot calcular un resultat quan la sèrie temporal original conté valors desconeguts

$$f(S, \tau, \delta) = m \text{ on } \exists n \in S : V(n) = \infty$$

Com ja hem comentat, treballar amb valors desconeguts estrictament hauria de resultar en valors desconeguts. Això no obstant, les dades desconegudes es poden tractar mitjançant tècniques de reconstrucció, d'interpolació, d'aproximació, etc. L'usuari, però, s'ha d'assegurar i estudiar en cada context que la tècnica que apliqui per a tractar dades desconegudes sigui vàlida. Altrament, només podrà considerar el resultat com a desconegut.

Per exemple, podem redefinir el patró de la funció d'agregació mitjana en una mitjana^u que sigui capaç de tractar valors desconeguts conservant l'àrea coneguda,

és a dir, l'àrea total coneguda quedarà escampada en l'interval de consolidació.

$$\text{mitjana}^u(S,\tau,\delta) = m \text{ on}$$
$$V(m) = \frac{1}{\delta}\int_{\delta}^{\tau+\delta} S^u(t)dt \text{ i}$$
$$S^u(t) = \begin{cases} 0 & \text{si } S^r(t) = \infty \\ S^r(t) & \text{altrament} \end{cases}$$

Marcatge de dades desconegudes. Una funció d'agregació d'atributs f que marqui dades desconegudes és aquella que pot retornar una mesura de valor indefinit com a resultat

$$f(S,\tau,\delta) = m \text{ on } V(m) \in \mathbb{R}^*$$

Per exemple, podem definir un patró de funció d'agregació d'atribut màxim que retorni valor desconegut si hi ha un a mesura amb el valor més gran que 2; és a dir establim un límit superior de 2, cosa que identifiquem amb L2.

$$\text{màxim}^{L2}(S,\tau,\delta) = m \text{ on}$$
$$m = \begin{cases} (T(n),\infty) & \text{si } \exists o \in S[\tau,\tau+\delta] : V(o) > 2 \\ n & \text{altrament} \end{cases} \text{ i } n = \text{màxim}(S,\tau,\delta])$$

Part III.

Consideracions i reflexions sobre els models

6. Introducció a consideracions sobre els models

Un cop definits els models de SGST i de SGSTM, podem fer algunes consideracions i reflexions sobre aquests models. Principalment, considerem els temes següents:

- Exposem SGSTM per a dispositius on l'emmagatzematge reduït i afitat és important. Aquest és bàsicament el model que hem presentat però en farem algunes consideracions més.
- Formulem una funció de multiresolució que permet expressar l'acció dels SGSTM com una funció sobre una sèrie temporal que retorna una nova sèrie temporal o un conjunt de sèries temporals.
- Exposem SGST amb emmagatzematges massius on calen consultes i visualitzacions ràpides computades mitjançant SGSTM.
- Formulem el problema de la qualitat en els SGSTM, introduïm el problema d'avaluar la informació que comprimeixen els SGSTM.

El primer tema correspon a petites variacions sobre el model definit. Ho presentem en una secció a continuació.

Els altres temes són consideracions i reflexions que usen el model definit com a referència. Les presentem cadascuna en un capítol diferent. Per als dos darrers també usem la formulació presentada en el segon tema.

6.1. Algunes variacions dels SGSTM

En el model de SGSTM hem definit l'estructura al més genèrica i senzilla possible per a encabir-hi diferents supòsits de multiresolució. Així doncs, es poden formular variacions dels SGSTM que en canviïn algun aspecte del comportament.

Particularment, en el model hem generalitzat els buffers de manera que s'acumula tota la sèrie temporal original independentment en cadascun dels buffers. Aquesta estructura és útil en el model perquè permet definir de forma molt abstracta el comportament dels SGSTM i abastar-ne diferents possibles variacions. Però en algunes implementacions del model pot ser útil utilitzar altres aproximacions, és a dir

estudiar algunes variacions podria resultar útils en el nivell físic on no es gaudeixen els avantatges abstractes matemàtics del nivell lògic.

De forma senzilla, podem pensar en implementacions que utilitzin els buffers d'una mateixa sèrie temporal de manera compartida, per exemple les diferents resolucions amb el mateix pas de consolidació poden compartir buffer. De forma més elaborada podem exposar implementacions del model en què s'emmagatzemi tota la sèrie temporal original en un SGST i el SGSTM treballi sobre aquestes mesures, és a dir que realment els buffers no les emmagatzemin sinó que seleccionin les mesures que necessiten a cada moment. En aquest cas pensem en SGST d'emmagatzematge massiu, com els descrits a l'apartat 2.3.4, i dels quals a la capítol 8 n'explorarem aplicacions mitjançant sistemes SGST i SGSTM conjunts.

A continuació considerem algunes petites variacions en els buffers i la consolidació que poden conduir cap a altres aplicacions. Presentem tres variacions de l'estructura:

- Resolucions encadenades
- Funcions d'agregació d'atributs orientades a flux
- El rellotge de consolidació

6.2. Resolucions encadenades

Una sèrie temporal multiresolució amb estructura de resolucions encadenades té la mateixa estructura que la presentada en el model (v. cap. 5) llevat que hi ha buffers que reben les mesures d'altres discs en comptes de l'entrada comuna de mesures. És a dir, que una subsèrie resolució depèn dels valors consolidats a una altra subsèrie resolució, cosa que anomenen resolucions encadenades.

La figura 6.1 mostra l'arquitectura d'una base de dades multiresolució ja presentada a la figura 5.1 però ara modificada amb les resolucions encadenades. En aquest cas, les mesures del disc de R_0 s'utilitzen en altres buffers i les mesures del buffer de R_k provenen d'un altre disc. En un cas simple de resolucions encadenades, podem considerar que el buffer d'una resolució és exactament el disc de l'altra. En un cas més elaborat, podem considerar que quan el disc d'una resolució descarta una mesura, s'afegeix al buffer de l'altra.

Respecte a l'estructura general, l'estructura encadenada restringeix els passos de consolidació dels buffers i els cardinals màxims dels discs. Els buffers que depenen d'una altra resolució han de tenir un pas de consolidació múltiple de l'altra resolució i han de tenir un període de buffer que estigui inclòs en el lapse de l'altra resolució. A més les resolucions encadenades també han de ser coherents en la funció d'agregació d'atributs, la qual pot ser que hagi de ser la mateixa funció. Les resolucions encadenades requereixen un estudi més profund que l'estructura general

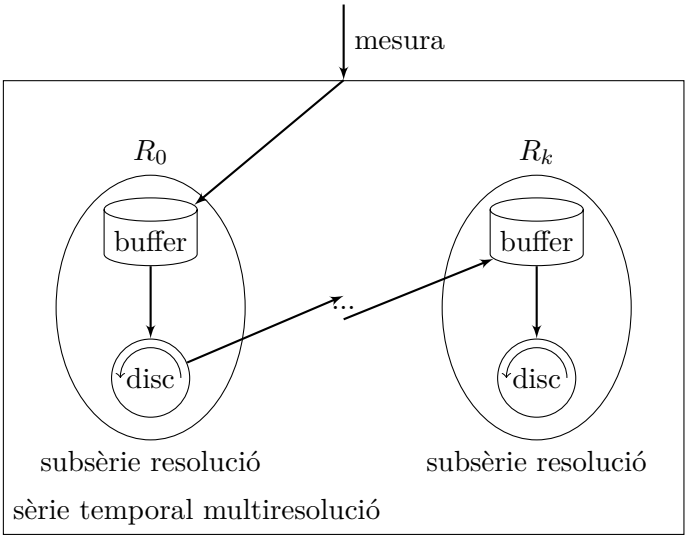


Figura 6.1.: Arquitectura encadenada d'una base de dades multiresolució

i poden encadenar pèrdues successives significatives, com per exemple és el cas de calcular la mitjana successivament que, per no ser associativa, no és el mateix que calcular-la en dos buffers independents.

L'estructura de resolucions encadenades pot ser útil per a aplicacions que necessitin distribuir l'emmagatzematge de les sèries temporals multiresolució. En l'estructura genèrica del model, cada mesura que s'insereix a una base de dades ha d'inserir-se a totes les subsèries resolució, és a dir que en cas d'un emmagatzematge distribuït tota la sèrie temporal original s'ha de distribuir a cada subsèrie. En canvi en l'estructura de resolucions encadenades, la sèrie temporal original primer es resumeix en una subsèrie resolució i és només aquest resum el que es distribueix a la següent subsèrie resolució. D'aquesta manera l'emmagatzematge de les resolucions queda distribuït en diferents nodes i a l'hora de respondre a una consulta només cal recollir les resolucions ja resumides que es necessitin. Deligiannakis, Kotidis i Roussopoulos [35] proposen una estratègia similar de disseminació de la informació per a xarxes de sensors.

A continuació mostrem, mitjançant un exemple, la variació que comporten les resolucions encadenades en el model de SGSTM.

Exemple 6.1 (Sèrie temporal multiresolució amb resolucions encadenades)**.** Per a definir una sèrie temporal multiresolució amb resolucions encadenades és útil reprendre l'exemple 5.2 en què s'exemplifica una sèrie temporal multiresolució organitzada en vistes.

En aquest cas la relació de sèries temporals i noms segueix sent la mateixa $L = ((S',$ $S), \{(S_{B1}, \{(26,0),(29,0)\}), (S_{D1}, \{(10,0),(15,0),(20,0),(25,0)\}), (S_{D2}, \{(10,0),(20,$

M

S'_B	S'_D	τ	δ	κ	f
S_{B1}	S_{D1}	25	5	4	mitjana
$\mathbf{S_{D1}}$	S_{D2}	20	10	3	mitjana

L

S'	S	
	t	v
S_{B1}	26	0
	29	0
S_{D1}	10	0
	15	0
	20	0
	25	0
S_{D2}	10	0
	20	0

Figura 6.2.: Taula d'una sèrie temporal multiresolució amb resolucions encadenades

0)})}) llevat que no hi ha S_{B2}, i la sèrie temporal multiresolució amb noms com a domini dels atributs de sèries temporals també és la mateixa $M = ((S'_B, S'_D, \tau, \delta, \kappa, f), \{(S_{B1}, S_{D1}, 25, 5, 4, \text{mitjana}), (\mathbf{S_{D1}}, S_{D2}, 20, 10, 3, \text{mitjana})\})$ excepte que el buffer de la segona resolució és el disc de la primera S_{D1}, el qual el destaquem en negreta. Mostrem L i M en forma de taula a la figura 6.2.

Se segueix aplicant la mateixa operació de vista M_2 que a l'exemple 5.2 per a obtenir la sèrie temporal multiresolució. A la figura 6.3 particularitzem l'arquitectura de la figura 6.1 per a la base de dades d'aquest exemple. Cal, però, tenir dues consideracions en les operacions estructurals dels SGSTM per a les resolucions encadenades:

- L'operació d'inserció de mesures, afegeixM(M, m), no pot inserir la mesura a tots els buffers de les subsèries resolució sinó només a aquells que no estiguin encadenats. En el cas de l'exemple només al buffer B_1. Aquests buffers, als quals podem anomenar buffers d'entrada, es poden expressar amb l'operació $\Pi_{\{S'_B\}}(M) - \rho_{S'_D/S'_B}\left(\Pi_{\{S'_D\}}(M)\right)$.
- Només es poden eliminar les mesures dels buffers que no siguin encadenats, és a dir dels buffers d'entrada. Les resolucions encadenades només poden llegir les dades dels altres discs però no hi tenen control.

6.3. Funcions d'agregació amb orientació a flux

Les funcions d'agregació d'atributs definides a la secció 5.3 operen sobre un interval de la sèrie temporal i retornen una mesura que en resumeix un atribut. Aquesta definició genèrica implica que els buffers han d'emmagatzemar temporalment un conjunt de mesures de la sèrie temporal original i un cop resumides les poden eliminar.

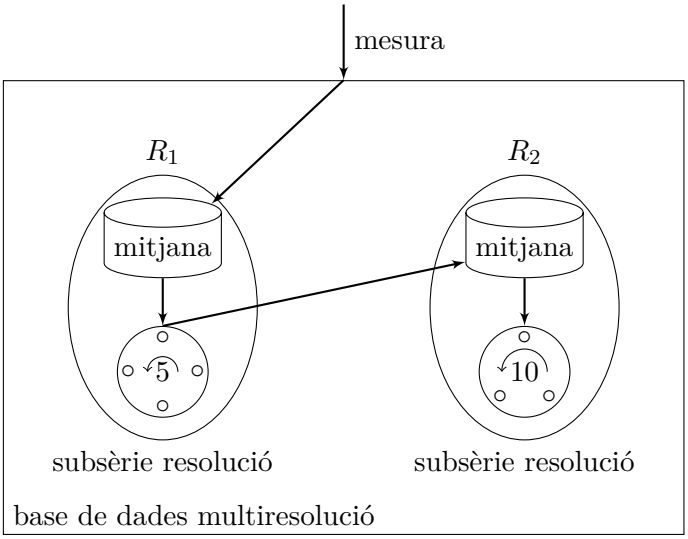


Figura 6.3.: Arquitectura de la base de dades multiresolució particular per l'exemple 6.1

Això no obstant, es poden utilitzar els algoritmes d'orientació a flux, com els que proposen Cormode, Korn i Tirthapura [18], per tal d'afitar els cardinals dels buffers. Tot i així no totes les funcions d'agregació d'atributs es poden implementar amb orientació a flux.

Definim una funció d'agregació amb orientació a flux com aquella que implementa el comportament equivalent a una funció d'agregació d'atributs, la qual hem notat anteriorment com a $f(S, \tau, \delta)$ on S és la sèrie temporal agregada i τ i δ són els paràmetres de consolidació. A diferència, però, una funció d'agregació amb orientació a flux, que notem com a flux, treballa sobre dues mesures $m' = \text{flux}(m, n, \tau, \delta)$ per a retornar la mesura resultant m', on n és la nova mesura que s'ha d'incorporar al flux, m és el flux anterior ja processat. Per a exemplificar-ho, redefinim les funcions d'agregació DD màxim i mitjana (v. def. 5.37) per tal que tinguin orientació a flux:

- $\text{flux\_màxim}^{\text{DD}}(m, n, \tau, \delta) = m'$ on $V(m') = \max(V(m), V(n))$ i $T(m') = \frac{2\tau+\delta}{2}$.
- $\text{flux\_mitjana}^{\text{DD}}(m, n, \tau, \delta) = m'$ on $V(m') = V(m) + \frac{V(n)}{\delta}$ i $T(m') = \frac{2\tau+\delta}{2}$.

Així, sigui $S = \{m_0, \ldots, m_k\}$ una sèrie temporal i τ i δ els paràmetres de consolidació. La funció d'agregació d'atributs $\text{mitjana}^{\text{DD}}$ calcula $m' = \text{mitjana}^{\text{DD}}(S, \tau, \delta)$. En canvi, la funció equivalent en flux $\text{flux\_mitjana}^{\text{DD}}$ calcula $m'_0 = \text{flux\_mitjana}^{\text{DD}}((0, 0), m_0, \tau, \delta), m'_1 = \text{flux\_mitjana}^{\text{DD}}(m'_0, m_1, \tau, \delta), \ldots, m'_k = \text{flux\_mitjana}^{\text{DD}}(m'_{k-1}, m_k, \tau, \delta)$, on $(0, 0)$ és una mesura inicial per al flux. Podem observar, doncs, que els dos càlculs són equivalents $m'_k = \text{mitjana}^{\text{DD}}(S, \tau, \delta)$.

Per a utilitzar en els SGSTM les funcions d'agregació d'atributs amb orientació a flux s'han de canviar els operadors d'afegir i de consolidar dels buffers:

- Sigui la definició 5.11, se'n modifica el comportament perquè $B = (m, \tau, \delta, f)$ sigui un buffer que emmagatzema una mesura m en comptes d'una sèrie temporal i l'operació d'afegir sigui afegeixB$(B, n) = (m', \tau, \delta, f)$ on $m' = f(m, n, \tau, \delta)$ i f és una funció d'agregació d'atributs orientada a flux.
- Sigui la definició 5.12, se'n modifica el comportament perquè essent el buffer modificat $B = (m, \tau, \delta, f)$ l'operació de consolidar sigui consolidaB$(B) = (B', m)$ on $B' = (m', \tau + \delta, \delta, f)$ i m' és l'element d'identitat de flux. Per exemple $m' = (0, 0)$ per a l'atribut de mitjana i $m' = (0, \min(\mathcal{V}))$ per a l'atribut de màxim on $\mathcal{V}$ és el domini dels valors.

Així doncs, en l'orientació a flux de les funcions d'agregació d'atributs, la mesura resultant es computa durant l'operació d'afegir noves mesures al buffer i quan s'ha de consolidar el buffer el resultat ja està disponible, només cal determinar l'element que actua com a identitat per a la funció d'agregació amb flux. En aquest cas, no té sentit parlar de l'eliminació de mesures antigues en el buffer.

6.4. Rellotge de consolidació

En el model de SGSTM no hi ha definit el concepte de rellotge, és a dir no s'explicita quan s'ha de computar l'operació de consolidar, si bé s'ha definit quan les sèries temporals multiresolució esdevenen consolidables. Les mesures tenen l'atribut de temps i, si s'insereixen ordenades, ja marquen el pas del temps. Tot i així, segons com sigui el rellotge i quan es computi l'operació de consolidació hi pot haver els escenaris següents:

- Extern. Ho anomenem rellotge extern o *push* perquè les mesures són les que controlen el procés de consolidació, de fet, el controla un sistema de monitoratge extern. El SGSTM no té rellotge sinó que s'utilitza l'atribut de temps de les mesures per conèixer l'instant actual. És el cas que hem definit en el model, en què una sèrie temporal multiresolució esdevé consolidable segons els instants de temps de les mesures adquirides i llavors ja pot ser consolidada. Per saber quan esdevé consolidable es pot consultar periòdicament o en base a esdeveniments, per exemple cada cop que s'insereixi una nova mesura. Ja que el temps observat pel SGSTM només canvia quan té mesures noves, això pot causar un cert decalatge de la consolidació de l'esquema amb un rellotge real, sobretot quan hi hagi inframostreig.
- Intern. Ho anomenem rellotge intern o *pull* perquè el SGSTM té un rellotge que controla el procés de consolidació. En aquests cas, la consolidació actua al marge del temps que indiquin les mesures i es computa quan ho marca

el rellotge. Això causa que la consolidació de l'esquema estigui totalment sincronitzada amb el rellotge real. En aquest cas s'hi poden incloure SGSTM que controlin el procés d'adquisició, és a dir que ordenin quan s'han d'adquirir noves mesures.

- Relatiu. Ho anomenem rellotge relatiu perquè el temps depèn d'altres esdeveniments. Per exemple que la consolidació es computi cada quatre mesures inserides. En aquest cas, es pot pensar en SGSTM i sistemes de monitoratge que no tinguin una bona mesura del temps actual, per exemple sense sincronització de rellotge, i en què l'objectiu dels SGSTM sigui informar de l'evolució de les variables situant-les relativament a partir de l'instant en què es fa la consulta. També pot ser el cas de les resolucions encadenades; com que depenen de la consolidació d'un altre disc aquest pot servir per a marcar el rellotge de consolidació dels buffers encadenats.

Així doncs, en les implementacions dels SGSTM cal decidir com ha de ser el rellotge de consolidació depenent del context on s'hagin d'aplicar. En els casos del model que hem definit, hem assumit un rellotge extern perquè permet definir exactament la consolidació a partir de la sèrie temporal original. En els altres dos tipus de rellotge, la consolidació esdevé variable i per tant la sèrie temporal resultant depèn, a més, del rellotge. És a dir, en aquests dos darrers casos els SGSTM participen activament en l'adquisició de les dades mentre que en el cas del rellotge extern els SGSTM són passius.

7. Funció de multiresolució aplicada a les sèries temporals

En aquest capítol definim la multiresolució com una funció que s'aplica a una sèrie temporal i retorna una nova sèrie temporal. Aquesta nova sèrie temporal és el resultat d'aplicar, en l'àmbit dels SGST, un esquema de multiresolució a la sèrie temporal original. Aquesta funció té el mateix efecte que utilitzar, en l'àmbit dels SGSTM, una sèrie temporal multiresolució amb el mateix esquema. Tot i així, les aplicacions que resulten d'ambdós casos no tenen les mateixes propietats.

En els capítols anteriors, hem definit la multiresolució com una estructura de base de dades per a emmagatzemar i tractar sèries temporals. La multiresolució també podria ser útil per a operar directament sobre una sèrie temporal, sense la capacitat d'emmagatzematge de dades. En aquest capítol avaluem com la funció de multiresolució por aplicar-se directament als SGST. Aquesta funció permet:

- Expressar problemes en què la multiresolució sigui gestionada com una consulta sobre una sèrie temporal; és a dir, com una operació dels SGST que calcula parelles d'agregació d'atributs i resolucions temporals per a una sèrie temporal de la mateixa manera com ho calculen els SGSTM (segons el model del capítol 5).
- Dissenyar sistemes duals de multiresolució, en els quals una sèrie temporal és emmagatzemada doblement en un SGSTM i en un SGST amb capacitats de resoldre funcions de multiresolució (v. cap. 8).
- Estudiar l'aplicació de la teoria de la informació per a determinar l'efecte que produeix la multiresolució en comprimir unes dades (v. cap. 9).
- Estudiar altres implementacions de la consulta de multiresolució, per exemple amb computació distribuïda i paral·lela (v. cap. 12).

A continuació descriurem com aplicar un esquema de multiresolució a una sèrie temporal mitjançant les operacions mapa i plec dels SGST. Això serà equivalent funcionalment als SGSTM.

7.1. Funció de multiresolució

En el model de SGSTM del capítol 5 hem definit un model de dades per a gestionar sèries temporals multiresolució. Aquest model té una estructura que emmagatzema la informació d'una sèrie temporal d'una forma determinada denominada multiresolució: en l'emmagatzematge és compacten les dades i es resumeix la informació per a consultes posteriors.

Per aquest motiu, el model de SGSTM té capacitats de computació sincronitzada o en línia (*online*) amb el temps i té característiques dels sistemes que tracten fluxos de dades (*data stream*); és a dir, dades que es van adquirint contínuament i gestionant al mateix temps. Això no treu, però, que de manera més simplificada també es pugui treballar amb un SGSTM en temps diferit (*offline*); és a dir, que primer s'emmagatzemin les dades adquirides i després, en temps posteriors, s'hi apliqui la consolidació.

A continuació, simplifiquem el càlcul de la multiresolució per a poder-lo aplicar, en temps diferit, directament als SGST. Aquest nou càlcul consisteix en una funció que transforma una sèrie temporal a una nova sèrie temporal.

7.2. Context de la formulació

Expressem el context de la funció de multiresolució que ens permet formular dues operacions mapa i plec amb un funcionament equivalent a un SGSTM.

Sigui $S = \{m_0, m_1, \ldots, m_k\}$ una sèrie temporal, M una sèrie temporal multiresolució i $E = \{(\delta_0, f_0, \tau_0, \kappa_0), \ldots, (\delta_d, f_d, \tau_d, \kappa_d)\}$ l'esquema de multiresolució de M (v. def. 5.10). Els tres passos, de forma resumida, per a calcular la multiresolució d'una sèrie temporal en un SGSTM són els següents.

1. S'afegeixen totes les mesures de la sèrie temporal S a la sèrie temporal multiresolució M, recursivament:

   $$M_0 = \text{afegeixM}(M, m_0), M_1 = \text{afegeixM}(M_0, m_1), \ldots, M_k = \text{afegeixM}(M_{k-1}, m_k)$$

2. Es consolida la sèrie temporal multiresolució resultant anterior, M_k, fins que no sigui consolidable (v. def. 5.13). Sigui $l \in \mathbb{N}$, recursivament

   $$M'_0 = \text{consolidaM}(M_k), M'_1 = \text{consolidaM}(M'_0), \ldots$$

   $$\ldots, M'_l = \text{consolidaM}(M'_{l-1}), M' = \text{consolidaM}(M'_l)$$

   on $M_k, M'_0, M'_1, \ldots, M'_{l-1}, M'_l$ són consolidables i M' no és consolidable.

3. Es consulta la sèrie temporal multiresolució amb les dues consultes bàsiques definides a l'apartat 5.2.3. Aquests dues consultes retornen les sèries temporals $S' = \text{SèrieTotal}(M')$ i $S'_{\delta f} = \text{SèrieDisc}(M', \delta, f)$; on δ i f són dos paràmetres qualssevol de l'esquema de multiresolució E, és a dir que existeix $(\delta, f, \tau, \kappa) \in E$.

En aquest context, formulem les funcions de transformació que permeten calcular en un SGST les sèries temporals S' i $S'_{\delta f}$ a partir de la sèrie temporal original S. Són dues funcions de multiresolució que anomenem mapa de multiresolució (MapMu) i plec de multiresolució (PlecMu). Així doncs, l'equivalència d'aquestes funcions amb les sèries temporals calculades en un SGSTM és la següent:

$$\text{MapMu}(S, \delta, f, \tau, \kappa) = S'_{\delta f}$$

$$\text{PlecMu}(S, E) = S'$$

és a dir amb un funcionament equivalent, en computació diferida, a

$$\text{SèrieDisc}(M', \delta, f) = \text{MapMu}(S, \delta, f, \tau, \kappa)$$
$$\text{SèrieTotal}(M') = \text{PlecMu}(S, E)$$

En resum. Primer, s'insereixen les mateixes mesures a un SGST i a un SGSTM. Després, en temps diferit: per una banda es consolida el SGSTM i es consulta la sèrie total i el conjunt de subsèries resolució emmagatzemades en els disc; per altra banda es calcula en el SGST l'operació de PlecMu i el conjunt d'operacions possibles de MapMu. Aleshores s'obtenen, respectivament, la mateixa sèrie temporal i el mateix conjunt de subsèries temporals.

7.3. Definicions

En primer lloc, definim l'operació de mapa mitjançant l'operador de mapa dels SGST (v. def. 4.31).

Definició 7.1 (Mapa de multiresolució)**.** *Sigui S una sèrie temporal i $(\delta, f, \tau, \kappa)$ un tuple de paràmetres d'un esquema multiresolució. El mapa de multiresolució és*

$$\text{MapMu}(S, \delta, f, \tau, \kappa) = \text{mapa}(R, g)$$

on

$$R = \{(t, \infty) | t \in Z\},$$
$$g(m) = f(S, T(m) - \delta, \delta)$$
$$Z = \{\tau + n\delta | n \in \mathbb{Z} \ \wedge \ T(\max S) - \kappa\delta < \tau + n\delta \leq T(\max S)\},$$

En segon lloc, definim l'operació de plec mitjançant l'operador de plec amb ordre dels SGST (v. def. 4.34) aplicat a l'esquema de multiresolució.

Definició 7.2 (Plec de multiresolució, amb comportament de SèrieTotal)**.** *Sigui S una sèrie temporal i $E = \{(\delta_0, f_0, \tau_0, \kappa_0), \ldots, (\delta_d, f_d, \tau_d, \kappa_d)\}$ un esquema de multiresolució. El plec de multiresolució d'una sèrie temporal és*

$$\mathrm{PlecMu}(S, E) = \mathrm{oplec}(E, \emptyset, g, \min)$$

on, usant la funció de la la definició 7.1,

$$g(R, (\delta, f, \tau, \kappa)) = R || \,\mathrm{MapMu}(S, \delta, f, \tau, \kappa).$$

Així, el plec de multiresolució és la concatenació de tots els MapMu possibles per a l'esquema E. De la mateixa manera que a la definició 5.35, s'ha d'assumir que E no conté δ repetits i que els concatenem per odre de δ. A més, noteu que en l'operació de plec tractem l'esquema E com una sèrie temporal multivaluada.

7.4. Exemples

Vegem en dos exemples el càlcul de la funció de multiresolució. Utilitzem la funció d'agregació d'atributs màxim$^{\mathrm{PD}}$ (v. def. 5.36).

Exemple 7.1 (Mapa de multiresolució)**.** Sigui la sèrie temporal $S = \{(1, 0), (3, 1), (6, 0), (10, 1)\}$ i els paràmetres de multiresolució ($\delta = 5, f = \text{màxim}^{\mathrm{PD}}, \tau = 0, \kappa = 2$). El mapa de multiresolució resulta en la sèrie temporal $S'_{5\,\text{màxim}^{\mathrm{PD}}} = \mathrm{MapMu}(S, 5, \text{màxim}^{\mathrm{PD}}, 0, 2)$ on $S'_{5\,\text{màxim}^{\mathrm{PD}}} = \{(5, 1), (10, 1)\}$. A continuació expressem aquest càlcul pas a pas i a la figura 7.1 es visualitzen en taula les sèries temporals corresponents:

1. El primer pas és obtenir els instants de temps que s'emmagatzemarien al disc d'una sèrie temporal multiresolució. Així, els instants de consolidació possibles són $Z' = \{\tau + n\delta | n \in \mathbb{Z}\} = \{\ldots, -5, 0, 5, 10, 15, \ldots\}$. Però un cop consolidat el disc només hi haurà els $\kappa = 2$ més recents abans de $T(\max S) = 10$, és a dir $Z = \{t \in Z' | T(\max S) - k\delta < t \leq T(\max S)\} = \{5, 10\}$.

2. El segon pas és obtenir a partir de Z la sèrie temporal R que correspon a la sèrie temporal que s'inicialitzaria al disc encara amb valors desconeguts, $R = \{(5, \infty), (10, \infty)\}$.

3. El tercer pas és calcular la funció d'agregació a S per a cada intervals de consolidació del disc de la forma $[T(m) - \delta, T(m)]$ on $m \in R$, és a dir $f(S, 0, 5)$ per a l'interval $[0, 5]$ i $f(S, 5, 5))$ per a l'interval $[5, 10]$. A tal efecte utilitzem el mapa sobre R per a calcular la sèrie temporal resultant $S'_{5\,\text{màxim}^{\mathrm{PD}}} = \{(5, f(S, 0, 5)), (10, f(S, 5, 5))\}$, en què $f = \text{màxim}^{\mathrm{PD}}$ i per tant resulta en els valors ja expressats $S'_{5\,\text{màxim}^{\mathrm{PD}}} = \{(5, 1), (10, 1)\}$.

S

t	v
1	0
3	1
6	0
10	1

R

t	v
5	∞
10	∞

B

t	v
5	t | v: 1 | 0; 3 | 1
10	t | v: 6 | 0; 10 | 1

$S'_{5\,\text{màxim}^{\text{PD}}}$

t	v
5	1
10	1

Figura 7.1.: Taules de les sèries temporals per l'operació MapMu

S

t	v
1	0
3	1
6	0
10	1

S_{D1}

t	v
5	1
10	1

S_{D2}

t	v
6	0
8	0
10	1

S'

t	v
5	1
6	0
8	0
10	1

Figura 7.2.: Taules de les sèries temporals per l'operació PlecMu

Es pot calcular un pas entremig per tal de mostrar les sèries temporals que hi hauria en el buffer abans de cada instant de consolidació. Així, per a cada $T(m)$ hi hauria la sèrie temporal $S[T(m) - \delta, T(m)]$, és a dir $B = \{(5, S[0, 5]), (10, S[5, 10])\} = \{(5, \{(1, 0), (3, 1)\}), (10, \{(5, 0), (10, 1)\})\}$.

Exemple 7.2 (Plec de multiresolució)**.** Sigui la sèrie temporal $S = \{(1, 0), (3, 1), (6, 0), (10, 1)\}$ i l'esquema de multiresolució $E = \{(\delta_0 = 5, f_0 = \text{màxim}^{\text{PD}}, \tau_0 = 0, \kappa_0 = 2), (\delta_1 = 2, f_1 = \text{màxim}^{\text{PD}}, \tau_1 = 0, \kappa_1 = 3)\}$. El plec de multiresolució resulta en la sèrie temporal $S' = \text{PlecMu}(S, E)$ on $S' = \{(5, 1), (6, 0), (8, 0), (10, 1)\}$. A continuació expressem aquest càlcul pas a pas I a la figura 7.2 es visualitzen en taula les sèries temporals corresponents:

1. En primer lloc es calcula la sèrie temporal pels paràmetres de multiresolució de δ_0: $S_{D0} = \text{MapMu}(5, \text{màxim}^{\text{PD}}, 0, 2) = \{(5, 1), (10, 1)\}$, com ja s'ha vist a l'exemple 7.1.

2. En segon lloc, es calcula la sèrie temporal pels paràmetres de multiresolució de δ_1: $S_{D1} = \text{MapMu}(2, \text{màxim}^{\text{PD}}, 0, 3) = \{(6, 0), (8, 0), (10, 1)\}$, de manera similar a com s'ha calculat S_{D0}.

3. En tercer lloc es concatenen les sèries temporals per ordre de δ: $\delta_1 < \delta_0$. Així, $S' = S_{D1} || S_{D0}$ que resulta en els valors ja expressats $S' = \{(5, 1), (6, 0), (8, 0), (10, 1)\}$.

8. Sistemes duals de multiresolució

Una sèrie temporal es pot emmagatzemar i gestionar en un SGST o en un SGSTM. També es pot dissenyar un sistema dual de multiresolució en què una sèrie temporal es tracti alhora en un SGST i en un SGSTM.

Les equivalències entre els SGSTM i les funcions de multiresolució aplicades a un SGST, formulades a la secció 7.1, permeten dissenyar sistemes duals que tinguin propietats complementàries. Així, aquests sistemes duals ofereixen altres utilitats a la multiresolució més enllà de l'orientació de compressió amb pèrdua que hem descrit en el model del capítol 5. A continuació:

- Dissenyem l'estructura d'aquests sistemes duals de multiresolució.
- Avaluem conceptes relacionats en l'àmbit genèric dels SGBD. Particularment la relació que hi ha amb la precomputació de consultes i amb el concepte de vistes dels SGBDR.
- Mostrem algunes aplicacions que permeten aquests sistemes duals: per a precomputar consultes, per a poder modificar els esquemes de multiresolució en un SGSTM, per a conservar totes les dades originals en un dipòsit massiu però que no cal consultar freqüentment, etc.

8.1. Estructura

Un sistema dual de multiresolució està format per un SGST i un SGSTM on s'emmagatzemen les mateixes sèries temporals. A cadascun s'hi poden fer les consultes pertinents de cada model per a les sèries temporal. A més, s'obté el mateix resultat en els dos sistemes per a les consultes que segueixin les restriccions de la funció de multiresolució formulades a la secció 7.1.

A la figura 8.1 es mostra l'estructura d'un sistema dual de multiresolució. L'usuari percep aquest sistema com un SGST on emmagatzema una sèrie temporal, S, i hi gestiona les consultes. Internament hi ha un SGST i un SGSTM que comparteixen l'entrada de mesures de la sèrie temporal. Així, quan l'usuari sol·licita una multiresolució, S', el sistema dual tant pot calcular-la a partir del SGST amb l'operació de $S' = \text{PlecMu}(S, E)$ com a partir del SGSTM amb l'operació de $S' = \text{SèrieTotal}(M)$, on E és un esquema de multiresolució i M és una sèrie multiresolució amb aquest

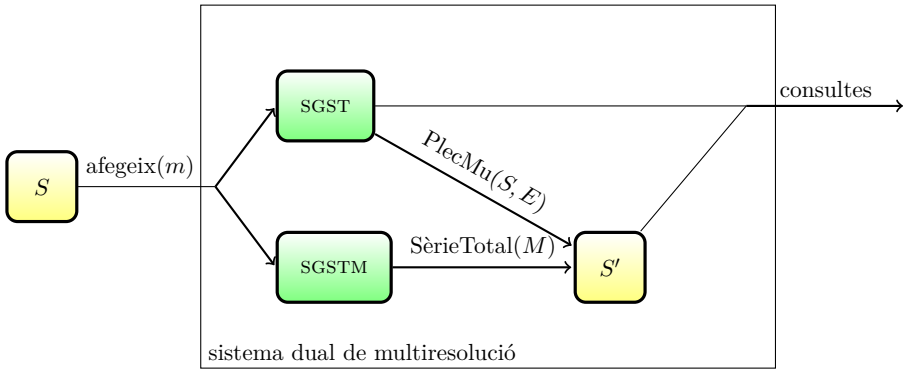


Figura 8.1.: Arquitectura dels sistemes duals de multiresolució: SGST+SGSTM

esquema. La mateixa estructura també pot servir per al cas de les operacions de MapMu i les de SèrieDisc.

Cal aclarir que el model de SGSTM està dissenyat en base al model de SGST i per tant aquests primers sempre depenen dels segons. No obstant això, cal no confondre aquesta dependència amb el sistema dual, el qual gestiona una mateixa sèrie temporal independentment en un SGSTM i en un SGST.

Tot i que per al sistema dual és equivalent calcular la sèrie temporal resultant a partir del SGST o del SGSTM, no pot seguir el mateix procediment en cada cas. Per una banda, la PlecMu(S) és una operació computada en temps diferit; cada cop que s'afegeix una nova mesura cal tornar a calcular tot el resultat. Per altra banda, la SèrieTotal(M) és una operació computada en línia; és a dir seguint el flux d'adicions de les mesures. Això no obstant, en un sistema de multiresolució dual no cal que les mesures s'emmagatzemin físicament en tots dos sistemes, sinó que els buffers dels SGSTM poden treballar amb les mesures emmagatzemades en un SGST massiu com hem comentat a la secció 6.1.

El sistema dual dissenyat funciona a partir de l'adició de mesures, de la mateixa manera que els SGSTM. L'ordre d'arribada d'aquestes mesures és crític en el sistema dual ja que, un cop el SGSTM s'ha consolidat, les dades més antigues que arribin no seran tingudes en compte i per tant l'equivalència entre les consultes de SGST i SGSTM ja no serà certa. Així doncs, si es vol mantenir l'equivalència, el sistema dual dissenyat té dues restriccions: només permet operacions d'adició i l'ordre d'adició és important. Més endavant, en les aplicacions d'aquests sistemes, descriurem l'abast d'aquestes restriccions.

8.2. Conceptes relacionats

En l'àmbit genèric dels SGBD, hi ha altres sistemes o altres conceptes semblants a l'estructura de sistema dual que proposem. Principalment s'utilitzen en dos àmbits similars: en la precomputació de consultes i en la precomputació de vistes.

Marz [91] i Marz i Warren [92] generalitzen un concepte similar al de sistema dual, ho emmarquen en l'àmbit dels SGBD per a *Big Data*. Proposen SGBD dissenyats amb tres nivells, que anomenen arquitectura *Lambda*:

- Nivell *batch*: Emmagatzema totes les dades originals i permet realitzar qualsevol consulta sobre aquestes dades. Preveu que algunes consultes operin sobre dades consultades prèviament, per tant en aquest nivell es gestionen també aquestes consultes precomputades, les quals a més es poden obtenir amb computació paral·lela com per exemple amb Hadoop. Es considera que les dades originals són immutables, és a dir que les bases de dades només permeten afegir però no modificar.
- Nivell *server*: Emmagatzema les consultes precomputades i n'ofereix les dades per a altres consultes. Les consultes precomputades s'han de tornar a calcular periòdicament i en el nivell *server* sempre hi ha la versió calculada més recent. Per tant, es preveu que les consultes precomputades no ofereixen la informació actualitzada al moment, sinó que hi ha un cert temps des que es modifiquen les dades originals fins que té impacte en les consultes.
- Nivell *speed*: Precomputa les mateixes consultes que el nivell *batch* però incrementalment, és a dir cada cop que s'afegeix una dada nova les dades de la consulta *speed* s'actualitzen adequadament. Aquest nivell només s'usa per a dades recents per tal de complementar el problema de les dades desactualitzades en els nivells *batch* i *server*.

L'arquitectura Lambda té moltes similituds amb el treball amb vistes en els SGBDR. Una vista (v. § 2.2.1) és un àlies per a una expressió relacional, és a dir una consulta que s'utilitza en altres consultes. Així doncs, una vista v és un àlies d'una consulta *op1* sobre unes *dades*, $v :=$ op1(dades), que si s'utilitza en una altra consulta, op2(v), és equivalent a executar totes dues operacions sobre les dades, op2(v) $\equiv$ op2(op1(dades)). En aquest sentit, el concepte de vista s'assembla a les consultes que es basen en altres consultes proposades per l'arquitectura Lambda de Marz i Warren i també a les sèries temporals multiresolució que proposem, les quals podem observar com a vista multiresolució d'una sèrie temporal original.

En el model relacional [23, cap. 10. Views] es considera, conceptualment, que les vistes no s'avaluen quan es defineixen sinó cada cop que s'executa una consulta una variable de la qual és una vista. En les implementacions, però, les vistes poden ser precomputades per tal d'emmagatzemar-ne temporalment els resultats i poder-los

reutilitzar en altres consultes; aleshores les vistes s'anomenen *snapshots* o *materialized views*. En el context de sistemes de suport a les decisions, la precomputació també es preveu en el càlcul de taules resum per a agregacions de les dades [23, cap. 22. Decision support]. Això no obstant, la precomputació de vistes no sempre comporta un millor rendiment; el concepte de vista del model permet la substitució algebraica i per tant també permet l'optimització global de la consulta i l'operació continguda a la vista.

Les vistes precomputades tenen associada una acció per actualitzar de nou el seu valor, és a dir per a recalcular la consulta que contenen quan les dades originals han canviat. En usar vistes precomputades cal preveure el termini de validesa dels càlculs precomputats, com ocorre en el nivell *server* de l'arquitectura *Lambda*. Així doncs, les vistes precomputades es poden actualitzar de vàries maneres:

- L'usuari decideix manualment quan s'han de tornar a computar. Per exemple, es pot utilitzar per a treballar amb les dades congelades a un cert instant en el temps (*snapshots*) sense haver de blocar les operacions de modificació de la base de dades [23, §10.5].
- Es computen periòdicament, com també es proposa en el nivell *batch* de l'arquitectura *Lambda*.
- Quan s'utilitzen per primer cop, es computen associades a un termini a partir del qual si es tornen a utilitzar s'hauran de tornar a computar.
- Es computen cada cop que es modifiquen les dades amb les quals operen, és a dir quan les dades originals reben una operació d'afegir, de modificar o d'actualitzar es torna a computar tota la vista.
- Es computen incrementalment. Quan es modifiquen les dades, s'aplica la mateixa operació a la vista precomputada; requereix, però, especificar quina operació s'ha d'aplicar per a actualitzar el resultat que ja hi ha a la vista precomputada. És a dir, inicialment es precomputa el resultat de la vista, sigui $v := \text{op1}(\text{dades})$. Quan es modifiquen les dades originals, amb una operació $\text{dades} := \text{op3}(\text{dades})$, s'ha de traslladar aquesta operació a la precomputació de la vista, amb una nova operació $v := \text{op3}'(v)$, de manera que s'aconsegueixi que $\text{op3}'(v) = \text{op3}(\text{op1}(\text{dades}))$. Així doncs, cal determinar $\text{op3}'$ as partir de op3, cosa que pot requerir un estudi complicat o fins i tot no ser possible. Aquesta translació és més senzilla quan només hi ha possibilitat d'operacions d'afegir noves dades però no de modificar-les: és el que es proposa en el nivell *speed* de l'arquitectura *Lambda* i el que admet el model de SGSTM que proposem. Jagadish, Mumick i Silberschatz [59] també proposen una solució semblant de mantenir vistes computades incrementalment segons arriben les dades amb un estudi contextualitzat en els *data stream*.

En resum, la sèrie temporal resultant dels sistemes duals de multiresolució pot ser considerada com una vista calculada sobre les sèries temporals originals. Aquesta

vista pot ser precomputada, cosa que en els sistemes duals de multiresolució es pot fer de dues maneres: mitjançant la funció de multiresolució en l'SGST, que s'ha de computar totalment cada cop que s'afegeix una nova mesura, i mitjançant l'SGSTM, que computa incrementalment. Aleshores aquestes vistes es poden usar per a altres consultes que tinguin com a context l'aproximació de multiresolució realitzada o per a visualitzacions gràfiques, com les que ofereix RRDtool [97].

8.3. Aplicacions

Les sèries temporals són dades que s'adquireixen contínuament i per tant cada cop és més gran el volum de dades que s'ha d'emmagatzemar i tractar. Aquest gran volum de dades és un problema per a operar amb les sèries temporals i és un problema en els sistemes que tenen l'emmagatzematge limitat. En aquest sentit, originalment hem plantejat el model de SGSTM per tal d'oferir una solució d'emmagatzematge que comprimeix la informació seleccionant-ne una multiresolució determinada.

Així, un SGSTM implica un selecció d'informació i la informació que no es considera importat és descartada. Aquests sistemes, per tant, no són adequats quan totes les dades monitorades han de ser emmagatzemades tal com s'adquireixen. Un cas d'aquests és quan no es coneixen quines funcions d'agregació són les més escaients per a les dades futures que s'adquiriran. Un altre cas és quan volem resoldre consultes detallades sobre les dades, com per exemple: a quina hora exacta ha ocorregut un esdeveniment.

Els sistemes duals de multiresolució ofereixen una solució per tal d'emmagatzemar totes les dades i alhora mantenen una gestió de multiresolució. En el sistema dual, s'ha d'entendre l'SGST com un emmagatzematge a llarg termini que no és consultat freqüentment; així pot estar implementat com a SGBD per a dades massives o basat en tècniques de compressió sense pèrdua. L'SGSTM s'ha d'entendre com un emmagatzematge de compressió amb pèrdua que conté multiresolucions precomputades de la sèries temporal. El temps de còmput no és tant crític en els SGSTM perquè es reparteix al llarg del temps, és a dir tal com es van adquirint les dades; més enllà del temps de còmput de cada funció d'agregació d'atributs, el qual limita la quantitat de multiresolucions diferents que pot gestionar un mateix SGSTM.

En la compressió de dades multimèdia s'utilitza una tècnica de gestió similar. Les dades s'emmagatzemen inicialment amb compressió sense pèrdua, a partir d'aquestes es generen dades amb compressió amb pèrdua que ocupen menys i són més àgils per a treballar. En el cas que calgui modificar les dades, es canvien les comprimides sense pèrdua i es regeneren de nou les comprimides amb pèrdua. Amb aquesta gestió s'evita el problema que les compressions amb pèrdua acumulin pèrdua entre successives modificacions (*generation loss*).

En resum, les aplicacions del sistema dual de multiresolució són les següents:

- Sistemes on els SGSTM precomputen incrementalment les consultes de multiresolució. És a dir, funcionen com a precomputació d'informació que es preveuen que es necessitarà; per tant al llarg del temps es creen i eliminen vistes segons les necessitats que es preveuen. Aleshores els SGST funcionen com a emmagatzematge a llarg termini que es consulta rarament. Aquesta aplicació és similar a la proposta de l'arquitectura Lambda i a la de les vistes precomputades incrementalment.
- Les dades emmagatzemades en els SGST s'utilitzen per al farciment inicial dels SGSTM gràcies a la funció de multiresolució que permet computar les sèries temporals dels discs a partir de l'operació MapMu. Pot tenir diversos objectius:
  - Quan es creen les vistes precomputades anteriors, inicialment el SGSTM contindrà sèries temporals amb valors desconeguts; amb la funció de multiresolució es poden inicialitzar amb els valors correctes.
  - Es pot usar per a canviar l'esquema de multiresolució dels SGSTM. En alguns canvis d'esquema, per exemple ampliar un disc, inicialment hi ha dades desconegudes però que es poden computar amb la funció de multiresolució.
  - Es pot usar per a canviar d'un emmagatzematge de sèries temporals en SGST a un emmagatzematge en SGSTM. Cal notar que és un canvi irreversible perquè l'emmagatzematge en els SGSTM és amb pèrdua.
- Es poden usar els SGST per a experimentar amb diversos esquemes de multiresolució per a les dades adquirides i així observar-ne la idoneïtat i escollir-ne un de millor.
- En el cas que no es compleixi la restricció d'ordre d'arribada de les mesures per a l'equivalència entre els SGST i els SGSTM, es podria refer la informació emmagatzemada en els SGSTM a partir dels SGST.

Com a contrapartida, però, en els sistemes duals apareix un SGST amb una gran quantitat de dades. Per tant, cal tenir en compte que si la informació computada pels SGSTM és suficient per a les consultes que s'han de realitzar, aleshores la informació emmagatzemada en els SGST és redundant. Això no obstant, no és senzill identificar i predir quan la informació emmagatzemada en el SGSTM serà totalment suficient; en el capítol 9 descrivim el problema d'identificar la informació que selecciona i la que perd un SGSTM.

En conclusió, encara que l'objectiu final sigui l'emmagatzematge de les sèries temporals comprimides amb pèrdua en un SGSTM, és a dir el model proposat originalment, els sistemes duals de multiresolució es poden utilitzar mentre hi hagi dubtes sobre quin esquema de multiresolució escollir i eliminar-los un cop es consideri que l'esquema és correcte. Aleshores, l'estructura de sistema dual serveix per a observar clarament que els SGSTM ofereixen una sèrie temporal que és una aproximació de

	SGST	SGSTM	SGST+SGSTM
càlcul de la multiresolució	S	S	S
emmagatzematge comprimit i limitat	N	S	N+S
pèrdua de dades	N	S	N
redundància de dades	N	N	S
actualització de la multiresolució	S	N	S
precomputacions en flux	N	S	S
precomputacions en paral·lel	S	N	S

Taula 8.1.: Comparació de les propietats dels SGST, dels SGSTM i dels duals SGST+SGSTM

l'original i que, per tant, permeten resoldre consultes aproximades. La taula 8.1 resumeix les propietats descrites dels SGST, dels SGSTM i dels duals SGST+SGSTM, on es comparen segons si les compleixen (S) o si no les compleixen (N).

9. Reflexions sobre la informació en la multiresolució

En aquest capítol reflexionem sobre el problema de la qualitat en la multiresolució de sèries temporals, és a dir sobre el problema d'identificar la informació que selecciona un esquema de multiresolució i per tant, alhora, d'observar quina informació no queda emmagatzemada i es perd. Contextualitzem aquest problema en l'aplicació de la teoria de la informació per a l'esquema de multiresolució.

En aquest context, els SGSTM són un sistemes que emmagatzemen dades comprimides amb pèrdua d'una certa part de la informació original. Aleshores, les consultes que es resolen a partir d'un SGSTM són consultes aproximades a la informació total original, llevat que mitjançant una anàlisi determinem que poden oferir consultes exactes. A continuació reflexionem sobre com analitzar l'error de les consultes de multiresolució respecte a la informació original:

- Primer, descrivim de forma molt genèrica la teoria de la informació, la qual està relacionada amb la quantificació de la informació.
- Segon, definim el problema de calcular l'error en la multiresolució.
- Tercer, mostrem exemples d'anàlisi de la informació per alguns esquemes de multiresolució.

9.1. Quantificació de la informació

En altres àmbits, la teoria de la informació (*information theory*) és la teoria de referència per a formalitzar la quantificació de la informació. Aquest també és el cas de la compressió de dades, un àmbit proper als SGSTM en aquest context d'informació. Per al cas particular de compressió amb pèrdua s'aplica un subconjunt de la teoria anomenat teoria de la taxa de bit-distorsió (*rate–distortion theory*); la qual modela la percepció de la distorsió i valora l'estètica dels resultats.

Per a quantificar la informació d'unes dades s'utilitza l'entropia, la qual mesura la incertesa que hi ha en unes dades particulars. Així, com més entropia més incertesa, a causa que les dades són més aleatòries, i com menys entropia més facilitat de predir-les a causa que són més redundants. Si l'entropia és zero aleshores les dades

són totalment predictibles; és a dir donat un valor es coneix exactament quin és el següent.

La compressió redueix la mida original de les dades. En el cas de la compressió sense pèrdua es conserva la informació però s'augmenta l'entropia, ja que s'eliminen les dades redundants. En el cas de la compressió amb pèrdua es descarta informació que es considera no essencial. Un exemple de compressió amb pèrdua és eliminar detalls d'una imatge que l'ull humà no pot apreciar. A més, la compressió amb pèrdua també es pot utilitzar per a transformar les dades a un altre domini on es percebi millor la informació, la qual cosa es coneix amb el nom de codificació perceptiva (*perceptual encoding*). Un exemple de codificació perceptiva és transformar els sons al domini freqüencial per a operacions d'equalització.

Els mètodes de compressió amb pèrdua se solen usar per a compressió de multimèdia. L'objectiu és aconseguir menys volum de dades però que conservin la mateixa percepció que les originals, o fins i tot amb una pèrdua de qualitat perceptible mentre compleixi amb els requisits de l'aplicació que se li vol donar. Així doncs, la compressió de multimèdia sovint requereix valorar la percepció humana, per tal de valorar la qualitat de percepció humana s'utilitzen testos subjectius en què un humà ha d'intentar distingir entre multimèdia original i multimèdia comprimit.

9.2. Error en la informació de la multiresolució

El model dels SGSTM es basa en una compressió amb pèrdua, és a dir descartar dades i emmagatzemar només aquella informació que es consideri necessària o suficient. Així doncs, cal quantificar quin error hi ha entre la informació emmagatzemada i la informació que contenen les dades originals.

El problema genèric de la informació en la multiresolució és el següent.

Definició 9.1 (Error en la informació de la multiresolució)**.** *Sigui una sèrie temporal S i una sèrie temporal multiresolució M resultant de l'emmagatzematge i consolidació de S. De la sèrie temporal multiresolució es pot consultar la sèrie temporal total $S' = \text{SèrieTotal}(M)$ o bé de forma equivalent, com s'ha descrit en el capítol 7, $S' = \text{PlecMu}(S, E)$ on E és l'esquema de M. S'executa una operació de consulta, o, sobre la sèrie temporal original, $r_1 = o(S)$, i la mateixa operació sobre la sèrie temporal de la multiresolució, $r_2 = o(S')$. Es pot avaluar l'error de multiresolució $\epsilon = d(r_1, r_2)$ on d és una funció que permet avaluar la distància entre els dos resultats i per tant considerem $\epsilon \geq 0$.*

Cal aclarir que es podria avaluar $d(S, S')$ com un problema d'aproximació al senyal original. Aquest és un problema ben resolt en altres models, com per exemple Last et al. [79] i Ogras i Ferhatosmanoglu [99], però no és el problema que es vol resoldre en els SGSTM. L'objectiu dels SGSTM és comprimir les dades i seleccionar una

determinada informació, per tant és un problema de compressió amb pèrdua. Així doncs, ens interessa avaluar l'error de la multiresolució en aquest context, en què les consultes als SGSTM (r_2) haurien de retornar resultats similars que si es fessin a les dades originals (r_1), sense que sigui necessari que S i S' es corresponguin.

Per a avaluar la distància $d(r_1, r_2)$, si els resultats d'ambdues consultes són sèries temporals, es pot utilitzar per exemple mínims quadrats com fa Last et al. [79]. Ara bé, en la informació de la multiresolució cal pensar també amb consultes qualitatives. Per exemple una consulta podria ser $o =$'Creix la sèrie temporal?' Aleshores no hi hauria error $\epsilon = d(r_1, r_2)$ quan la resposta fos la mateixa per als dos casos, r_1 i r_2, i hi hauria error quan les respostes diferissin.

Tot i que hem proposat d'aplicar la mateixa operació o a la sèrie temporal original $r_1 = o(S)$ i a la sèrie temporal de la multiresolució $r_2 = o(S')$, pot ser que l'operació hagi de ser diferent per a obtenir el mateix resultat. És a dir $r_1 = o(S)$ però $r_2 = o'(S')$ on o' és l'operació equivalent a o que s'ha d'aplicar després de la multiresolució. Per exemple, una operació o podria ser el càlcul de la multiresolució, $r_1 = \text{PlecMu}(S, E)$, aleshores l'operació o' equivalent és la identitat perquè el SGSTM ja ha calculat la multiresolució, $r_2 = S'$. Aquest exemple, de fet, és el cas que hem formulat a la secció 7.1, on hem descrit l'equivalència entre la sèrie temporal total d'un SGSTM i la funció PlecMu d'una sèrie temporal; per tant l'error $\epsilon = d(r_1, r_2)$ seria nul.

Així doncs, l'error en la informació de la multiresolució permet quantificar la pèrdua d'informació dels SGSTM. Un cop quantificat l'error per a un determinat esquema de multiresolució es pot saber per a quines consultes serà apropiat aquell esquema i per a quines no. Per a quantificar aquest error es poden preveure diversos contextos:

- Hi ha consultes que es poden resoldre a la perfecció, és a dir sense error. Per exemple és el cas descrit on l'operació que es vol fer a una sèrie temporal és precisament la consulta de multiresolució.
- Es pot calcular l'error mesurant la diferència entre la mateixa consulta aplicada a les dades originals que a les dades emmagatzemades en el SGSTM.
- Es pot avaluar subjectivament l'error mitjançant la visualització de les dades. És a dir, de manera semblant a la compressió amb pèrdua de multimèdia, l'usuari valora subjectivament si visualitza la mateixa informació en les dades originals com en les dades comprimides amb multiresolució. Per exemple, un dels criteris que recomana RRDtool per a establir un esquema de multiresolució és tenir en compte l'amplada de la pantalla on es visualitzen els resultats: no cal treballar amb una sèrie temporal amb molta resolució si no és possible observar-la [8, Rates, normalizing and consolidating] .

9.3. Exemples d'anàlisi de la informació

La definició 9.1 descriu el problema d'informació en la multiresolució de forma genèrica i abstracta. Així, en cada context d'aplicació de la multiresolució cal interpretar-ne el significat i particularitzar-ne un mètode d'anàlisi adequat. A tal efecte, a continuació proposem alguns exemples que mostren casos particulars d'anàlisi de la selecció o pèrdua d'informació que hi ha en un SGSTM.

Per a simplificar els exemples, proposem esquemes de multiresolució amb només una resolució i amb funcions d'agregacions d'atributs que seleccionen intervals independents de mesures. Per tal de referir-nos amb comoditat als càlculs de les funcions d'agregació d'atributs, anomenem les mesures i els valors amb els quals operen mitjançant la notació següent.

Definició 9.2 (Notació de les mesures i els valors amb els quals operen les funcions d'agregació d'atributs)**.** *Sigui $S = \{m_0, \ldots, m_k\}$ una sèrie temporal, on recordem que les mesures tenen la forma $m = (t, v)$, i $E = \{(\delta, f, \tau, \kappa)\}$ un esquema de multiresolució amb una sola subsèrie resolució. Assumim $\kappa = \infty$ per tal de negligir-ne l'efecte i $\tau = T(\min S)$. Obtenim la sèrie temporal resultant de la funció de multiresolució $S' = \mathrm{PlecMu}(S, E)$ (v. def. 7.2). Així, aquesta sèrie temporal resultant conté mesures calculades a partir de l'agregació de S en els intervals definits per τ i δ i té la forma*

$$\begin{aligned} S' = \{ & f(S, \tau, \delta), \\ & f(S, \tau + \delta, \delta), \\ & \ldots, \\ & f(S, \tau + (n-1)\delta, \delta)\} \end{aligned}$$

on $\tau + n\delta \geq T(\max S)$ i $n \in \mathbb{N}$.

La funció d'agregació d'atributs f realitza una operació sobre les mesures corresponents a l'interval de temps definit (v. § 5.3). Així per a l'agregació en el primer interval $[\tau, \tau + \delta]$ escriurem de forma genèrica que utilitza les mesures $m_0, \ldots, m_{i_1}$ de la sèrie temporal original, $m_0, \ldots, m_{i_1} \in S$ on i_1 pot ser qualsevol índex, i per tant escriurem els valors corresponent a aquestes mesures com $v_0, \ldots, v_{i_1}$. De la mateixa manera, escriurem $m_{i_1+1}, \ldots, m_{i_2}$ i $v_{i_1+1}, \ldots, v_{i_2}$ per a l'agregació en el segon interval, $[\tau + \delta, \tau + 2\delta]$, i així successivament fins al darrer interval en què notem les mesures amb $m_{i_n}, \ldots, m_k$ i els valors amb $v_{i_n}, \ldots, v_k$.

Aleshores, a partir de la notació dels valors, podem expressar la sèrie temporal resultant amb la forma

$$\begin{aligned} S' = \{ & (\tau + \delta, f'(v_0, \ldots, v_{i_1})), \\ & (\tau + 2\delta, f'(v_{i_1+1}, \ldots, v_{i_2})), \\ & \ldots, \\ & ((\tau + n\delta, f'(v_{i_M}, \ldots, v_k))\} \end{aligned}$$

on f' és l'operació corresponent de l'atribut que resumeix f. En els exemples següents assenyalarem quin f' correspon a cada f que utilitzem.

L'anàlisi que formulem és una introducció a la reflexió sobre l'error en la informació de la multiresolució. Així, de forma simple, analitzem si hi ha error o si no n'hi ha, sense pretendre avaluar quantificacions més complicades. A més, ho analitzem en base a l'esquema de multiresolució que s'utilitzi, particularment de quines funcions d'agregació d'atributs s'utilitzin i de com siguin les consultes posteriors.

9.3.1. Mateixa consulta i funció d'agregació d'atributs

En aquest apartat es formulen exemples que reflexionen sobre l'error de multiresolució que hi pot haver quan una consulta s'aplica a tota la sèrie temporal i l'operació de consulta es correspon amb la mateixa funció d'agregació d'atributs que s'ha utilitzat en l'esquema de multiresolució.

El context del problema és el següent. S'aplica un operador de consulta o a les dues sèries temporals, $r_1 = o(S)$ i $r_2 = o(S')$. Aquest operador o és el mateix càlcul que fa la funció d'agregació d'atributs f però aplicat a totes les mesures de la sèrie temporal original, $o(S) = V(f(S, [-\infty, \infty]))$. Analitzem l'error de multiresolució entre r_1 i r_2 segons tres funcions d'agregació d'atributs:

- Màxim: $f = \text{màxim}^{\text{PD}}$, el qual es correspon a aplicar l'operació d'agregació dels SGST $o = \text{max\_v}$ (v. def. 5.36) i per tant a calcular l'atribut $f' = \max$ dels valors (v. ex. 4.14).

  D'una banda $S = \{m_0, \ldots, m_k\}$ i aleshores $r_1 = o(S) = \text{max\_v}(S) = \max(v_0, \ldots, v_k)$. D'altra banda, aplicant la notació de la la definició 9.2, $S' = \{(\tau + \delta, \max(v_0, \ldots, v_{i_1})), (\tau + 2\delta, \max(v_{i_1+1}, \ldots, v_{i_2})), \ldots, (\tau + n\delta, \max(v_{i_n}, \ldots, v_k))\}$ i aleshores $r_2 = o(S') = \max\big(\max(v_0, \ldots, v_{i_1}), \max(v_{i_1+1}, \ldots, v_{i_2}), \ldots, \max(v_{i_n}, \ldots v_k)\big)$.

  Atès que $\max(v_0, \ldots, v_k) = \max\big(\max(v_0, \ldots, v_{i_1}), \max(v_{i_1+1}, \ldots, v_{i_2}), \ldots, \max(v_{i_n}, \ldots v_k)\big)$ perquè max és una funció associativa: $\max(a, b, c, d, e) = \max(\max(a, b), \max(c, d, e))$, podem concloure que en aquest cas $r_1 = r_2$ i per tant $\epsilon = 0$.

  En aquest exemple hem negligit els temps resultants de l'agregació. És a dir, r_1 és correspon amb una o més d'una mesura $m \in S : V(m) = r_1$ i de la mateixa manera $n \in S' : V(n) = r_2$ on hem conclòs que $V(m) = V(n)$. Això no obstant, en general els temps d'aquestes mesures no es correspondran, $T(m) \neq T(n)$ perquè $f = \text{màxim}^{\text{PD}}$ resumeix els atributs de temps segons l'interval de consolidació i al marge del resum de la informació en els valors.

- Mitjana aritmètica: $f = \text{mitjana}^{\text{PD}}$, el qual es correspon a aplicar l'operació d'agregació dels SGST $o = \text{mitjana\_v}$ (v. def. 5.36) i per tant a calcular l'atribut $f' = \text{mitjana}$ (aritmètica) dels valors (v. ex. 4.14).

  De manera similar al cas anterior, els resultats són $r_1 = \text{mitjana}(v_0, \ldots, v_k)$ i $r_2 = \text{mitjana}\big(\text{mitjana}(v_0, \ldots, v_{i_1}), \text{mitjana}(v_{i_1+1}, \ldots, v_{i_2}), \ldots, \text{mitjana}(v_{i_n}, \ldots v_k)\big)$. Però en aquest cas hem de concloure que $\epsilon \geq 0$ perquè la mitjana no és una funció associativa: $\text{mitjana}(a, b, c, d, e) \neq \text{mitjana}(\text{mitjana}(a, b), \text{mitjana}(c, d, e))$.

- Total: definim una funció d'agregació d'atributs $f = \text{total}$ que, negligint l'atribut de temps, resumeix la sèrie temporal amb la suma dels valors: $\text{total}(S, \tau, \delta) = m'$ i $V(m') = \sum_{\forall m \in S(\tau, \tau+\delta]} V(m)$. Així doncs, es correspon a aplicar l'operació d'agregació dels SGST $o = \text{suma\_v}$ i per tant a calcular l'atribut $f' = \sum$ dels valors (v. ex. 4.14).

  Aquest cas és similar al del màxim, $r_1 = \sum(v_0, \ldots, v_k)$ i $r_2 = \sum\big(\sum(v_0, \ldots, v_{i_1}), \sum(v_{i_1+1}, \ldots, v_{i_2}), \ldots, \sum(v_{i_n}, \ldots v_k)\big)$, i $\epsilon = 0$ perquè la suma és una funció associativa.

9.3.2. Mitjana d'una sèrie temporal regular

Seguint l'apartat 9.3.1, podem estudiar en quins casos la mitjana aritmètica no té error. Com ja s'ha dit, la mitjana no és una funció associativa. Però sí que esdevé associativa quan s'associen el mateix nombre d'elements: $\text{mitjana}(a, b, c, d, e, f) = \text{mitjana}(\text{mitjana}(a, b), \text{mitjana}(c, d), \text{mitjana}(e, f)) = \text{mitjana}(\text{mitjana}(a, b, c), \text{mitjana}(d, e, f))$.

Per tal que s'associïn el mateix nombre d'elements cal que per cada interval de consolidació de la sèrie temporal hi hagi el mateix nombre de mesures: $|S[\tau, \tau+\delta]| = |S[\tau + \delta, \tau + 2\delta]| = \cdots = |S[\tau + (n-1)\delta, \tau + n\delta]|$. Aquest cas es compleix, per exemple, quan la sèrie temporal és regular amb període p i el pas de consolidació de l'esquema multiresolució és múltiple de la regularitat de la sèrie temporal, $\delta = kp$ on $k \in \mathbb{N}$, o bé quan la sèrie temporal és de temps real amb període p i iniciada a t i la multiresolució n'és múltiple: $\delta = kp$ i $\tau = t + k\delta$ (v. § 4.3.3).

Aleshores, en aquest casos, sí que es podria concloure que $\epsilon = 0$ per la multiresolució amb mitjana.

9.3.3. Consulta d'un interval determinat

En l'apartat 9.3.1 l'operació de consulta o s'aplica a tota la sèrie temporal $S[-\infty, \infty]$. Ara proposem d'aplicar-la a un interval concret de la sèrie temporal $S[s, t]$ on s

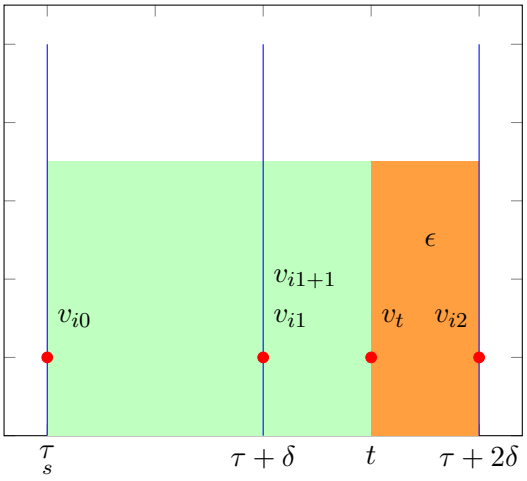


Figura 9.1.: Sèrie temporal amb la consulta desitjada (verd) i l'error de la informació no coneguda (taronja)

i t són dos instants de temps. Analitzem l'error de multiresolució quan l'interval $[s, t]$ és múltiple dels intervals de multiresolució consolidats, $s = \tau + k\delta$ i $t = \tau + (k+l)\delta$ on $k, l \in \mathbb{N}$, i quan no ho és.

L'interval és múltiple dels intervals de multiresolució consolidats. Aleshores $r_1 = V(f(S, [s, t]))$ i $r_2 = f(S', [s, t])$ on $S' = \{\ldots, (s+\delta, f'(v_s, \ldots, v_{s+1})), \ldots, (t, f'(v_{t-1}, \ldots, v_t)), \ldots\}$. En aquest darrer cas, només assenyalem els valors que se seleccionen, $S'[s, t]$, els quals notem amb $v_s, \ldots, v_{s+\delta}$ pels valors de les mesures en $S[s, s+\delta]$ i amb $v_{t-\delta}, \ldots, v_t$ pels valor en $S[t-\delta, t]$.

Per tant, per a estudiar $d(r_1, r_2)$ es pot analitzar el comportament de la funció de resum de l'atribut per als dos casos $r_1 = f'(v_s, \ldots, v_{s+\delta}, \ldots, v_{t-\delta}, \ldots, v_t)$ i $r_2 = f'(f'(v_s, \ldots, v_{s+\delta}), \ldots, f'(v_{t-\delta}, \ldots, v_t))$, cosa que és una situació similar a la de l'apartat 9.3.1.

L'interval no és múltiple dels intervals de multiresolució consolidats. Per a simplificar la notació, suposem $s = \tau$ i $\tau + \delta < t \leq \tau + 2\delta$. Aleshores $r_1 = V(f(S, [s, t]))$ i $r_2 = f(S', [s, t])$ on $S' = \{(\tau+\delta, f'(v_{i_0}, \ldots, v_{i_1})), (\tau+2\delta, f'(v_{i_1+1}, \ldots, v_t, \ldots, v_{i_2})), \ldots\}$. Els valors s'anoten com a la definició 9.2 però s'afegeix el valor v_t que assenyala una possible mesura en l'instant t. A la figura 9.1 es mostren els instants de temps i els valors d'aquest exemple, l'interval temporal $[s, t]$ de la consulta i l'interval temporal $[t, \tau + 2\delta]$ que mostra l'error en la consulta a partir de la informació emmagatzemada a la multiresolució.

En aquest cas, cal tenir en compte que per a calcular $f(S', [s, t])$ s'ha de resoldre la selecció de S' en l'interval $[s, t]$, cosa que es pot realitzar:

1. amb una selecció d'interval, $S'[s,t] = \{(\tau+\delta, f'(v_{i_0},\ldots,v_{i_1}))\}$,
2. amb una selecció d'interval temporal, $S'[s,t]^r = (\tau+\delta, f'(v_{i_0},\ldots,v_{i_1}),(t, f^r(f'(v_{i_1+1},\ldots,v_t,\ldots,v_{i_2}))))$ on f^r és la interpolació realitzada per la funció de representació r
3. o amb altres casos, podem pensar per exemple amb una funció de representació que directament utilitzi el valor de tot el segon interval com a vàlid per a t, $S'[s,t]^r = (\tau+\delta, f'(v_{i_0},\ldots,v_{i_1}),(t, f'(v_{i_1+1},\ldots,v_t,\ldots,v_{i_2})))$ on f^r no hi és perquè seria la funció identitat.

Així, per a estudiar $\epsilon = d(r_1,r_2)$ s'ha d'analitzar d'una banda $r_1 = f'(v_{i_0},\ldots,v_{i_1}, v_{i_1+1},\ldots,v_t)$ i de l'altra, depenent de quina de les tres seleccions s'utilitzi,

1. $r_2 = f'(f'(v_{i_0},\ldots,v_{i_1}))$,
2. $r_2 = f'(f'(v_{i_0},\ldots,v_{i_1}), f^r(f'(v_{i_1+1},\ldots,v_t,\ldots,v_{i_2})))$
3. o $r_2 = f'(f'(v_{i_0},\ldots,v_{i_1}), f'(v_{i_1+1},\ldots,v_t,\ldots,v_{i_2}))$.

És a dir, en la multiresolució consolidada no hi ha disponible la informació $f'(v_{i_1+1}, \ldots, v_t)$ que és la que es voldria consultar. Per tant hem de concloure que en aquest cas generalment $\epsilon = d(r_1,r_2) \geq 0$. Tot i així, en la segona selecció proposada es pot observar que si es coneix exactament el comportament de la sèrie temporal, per exemple s'estudia mitjançant la teoria del senyal, aleshores pot ser possible de determinar una funció de representació que compleixi $f'(v_{i_1+1},\ldots,v_t) = f^r(f'(v_{i_1+1}, \ldots, v_t,\ldots,v_{i_2}))$ i per tant aconseguir $r_2 = f'(f'(v_{i_0},\ldots,v_{i_1}), f'(v_{i_1+1},\ldots,v_t))$; cosa que ja és una situació similar a la de l'apartat 9.3.1..

Retornant, però, al cas que hi ha error $\epsilon = d(r_1,r_2) \geq 0$, estudiem dos dels atributs descrits a l'apartat 9.3.1 per tal d'avaluar si és possible afitar-ne l'error en aquests casos. Així, formulem el cas que caldria calcular $f'(v_{i_1+1},\ldots,v_t)$ però la informació que hi ha emmagatzemada per a aquest interval és $f'(v_{i_1+1},\ldots,v_t,\ldots,v_{i_2})$:

- Màxim. Cal calcular $\max(v_{i_1+1},\ldots,v_t)$ però hi ha emmagatzemat $\max(v_{i_1+1}, \ldots, v_t,\ldots,v_{i_2})$. Per tant l'error en la consulta és $\epsilon = d(r_1,r_2) = d(\max(v_{i_1+1}, \ldots, v_t), \max(v_{i_1+1},\ldots,v_t,\ldots,v_{i_2}))$. Si el màxim es troba a $[s+\delta,t]$ l'error és nul però si hi ha un màxim a $[t,s+2\delta]$ aleshores l'error és $\epsilon = d(\max(v_{i_1+1}, \ldots, v_t), \max(v_t,\ldots,v_{i_2}))$, el qual no és fitable perquè generalment no podem trobar cap relació entre $\max(v_{i_1+1},\ldots,v_t)$ i $\max(v_t,\ldots,v_{i_2})$.
- Total: Cal calcular $\sum(v_{i_1+1},\ldots,v_t)$ però hi ha emmagatzemat $\sum(v_{i_1+1},\ldots, v_t,\ldots,v_{i_2})$. Per tant l'error en la consulta és $\epsilon = d(r_1,r_2) = d(\sum(v_{i_1+1},\ldots, v_t,\ldots,v_{i_2}), \sum(v_{i_1+1},\ldots,v_t)) = \sum(v_{t+1},\ldots,v_{i_2})$. Però $\sum(v_{t+1},\ldots,v_{i_2})$ no és un valor emmagatzemat a la multiresolució i per tant no es pot saber l'error. Això no obstant, en el cas del total té sentit plantejar el cas que la variable mesurada és monòtona creixent (v. a continuació l'apartat 9.3.4): aleshores es compleix que $\sum(v_{i_1+1},\ldots,v_t) \leq \sum(v_{i_1+1},\ldots,v_t,\ldots,v_{i_2})$ i per tant es pot

fitar l'error $\epsilon = d(r_1, r_2) = \sum(v_{t+1}, \ldots, v_{i_2}) \leq \sum(v_{i_1+1}, \ldots, v_t, \ldots, v_{i_2})$; és a dir com a màxim es cometria un error del valor consolidat a $\tau + 2\delta$ que significaria que en els pitjors del casos tota la mesura hauria ocorregut després de t.

9.3.4. Conservació d'informació en comptadors

Els comptadors són un dels trets semàntics que poden tenir les sèries temporals (v. § 4.3.1) i se'n pot conservar la informació aprofitant aquests trets. En aquest cas explorem com conservar la mitjana de la funció dels comptadors. Aquest exemple prové d'una reflexió acurada de per què RRDtool té com a referent els comptadors.

Un comptador monòton creixent és un aparell que mesura l'energia en un determinat interval de temps. Entre dues lectures successives del comptador la mesura de l'energia és exacta a diferència d'un aparell que mesuri potència instantània. D'aquesta només es pot deduir l'energia exacta si es considera que el senyal es pot reconstruir, per exemple compleix la freqüència del teorema de Nyquist–Shannon tot i que a la pràctica és complicat conèixer la freqüència de les variables mesurades atès que solen ser aleatòries o canvien bruscament. A la inversa també ocorre el mateix, a partir de la mesura de l'energia només es pot deduir la potència instantània exacta si es considera que el senyal es pot reconstruir.

Els conceptes d'energia i potència solen anar associats a un determinat tipus de variables físiques contínues; en altres comptadors els conceptes equivalents són quantitat, comptatge total o increments per a l'energia i el flux o la velocitat per a la potència. No totes les variables físiques són susceptibles de ser mesurades amb un comptador. Els comptadors es poden aplicar per exemple per a mesurar energia elèctrica (v. ex. 4.19), aforaments de trànsit en una carretera, consum d'aigua, etc.

En resum, l'aparell condiciona la informació que es podrà extreure de la mesura; en aquest exemple ens centrem en la informació de l'energia i de quina manera la multiresolució és capaç de conservar exactament algunes propietats d'aquesta informació. Per a aquest exemple suposem aparells de mesura ideals quan parlem de mesura exacta; és a dir que no tenim en compte l'error de precisió o d'exactitud de l'aparell.

Així doncs, la definició del problema és la següent. Sigui $E(t)$ l'energia d'un senyal i $P(t)$ la potència instantània del senyal, es compleix la relació $E(t) = \int P(t)dt$, la qual es mostra a la figura 9.2. Siguin $[x, y]$ i $[y, z]$ dos intervals de temps de mesura, un comptador mesura exactament el valor de $E_s^t = \int_x^y P(t)dt$ i de $E_y^z = \int_y^z P(t)dt$. En canvi, un aparell de mesura de potència instantània es capaç de mesurar exactament $P(x)$, $P(y)$ i $P(z)$. Ara bé, a partir del comptador no es poden deduir exactament $P(x)$, $P(y)$ ni $P(z)$ i a partir de la mesura de la potència

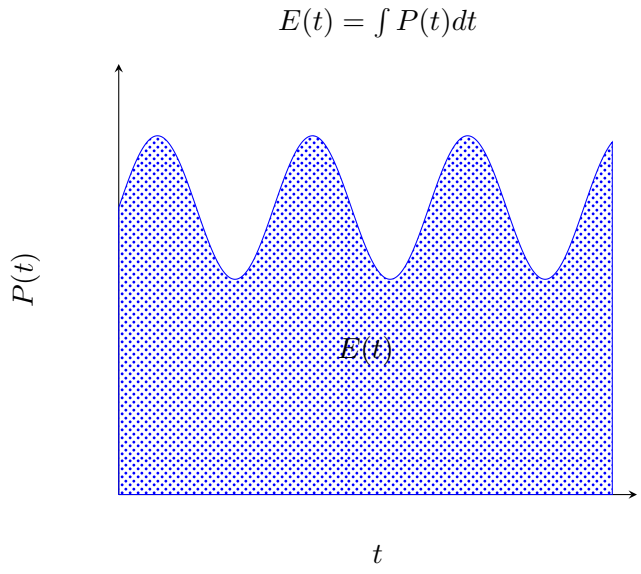


Figura 9.2.: Relació entre l'energia i la potència o la quantitat comptada i la velocitat

instantània no es poden deduir exactament $\int_x^y P(t)$ ni $\int_y^z P(t)$. Tampoc a partir del comptador es poden deduir exactament energies que no s'han mesurat, per exemple ni $\int_{(x+y)/2}^y P(t)$ ni $\int_{(x+y)/2}^z P(t)$; ara bé sí que serà exacte el càlcul $\int_x^y P(t) + \int_y^z P(t)$.

La multiresolució és capaç de conservar aquesta exactitud del comptatge total. El comptatge total es pot conservar en un esquema de multiresolució amb funcions d'agregació per atributs de suma de totals o bé per atributs de mitjana de la funció. Aquests darrers són els que permeten, a més, expressar la sèrie temporal resultant de la multiresolució de forma més coherent amb l'original (v. § 4.3.1). Reprenent l'apartat 9.3.1, avaluem els atributs de mitjana de la funció, els quals són semblants als atributs de total però considerant la sèrie temporal en la representació contínua:

- Mitjana de la funció: f = mitjana, segons el patró general de mitjana de la funció mostrat a l'apartat 5.3.1, el qual es correspon a calcular la mitjana de la funció de representació de la sèrie temporal $\frac{1}{b-a}\int_a^b S(t)dt$ en l'interval tancat $[a, b]$.

  Sigui $t_M = T(\max(S))$ i $t_m = T(\min(S))$ i $S' = \{(\tau+\delta, \frac{1}{\delta}\int_\tau^{\tau+\delta} S(t)dt), (\tau+2\delta, \frac{1}{\delta}\int_{\tau+\delta}^{\tau+2\delta} S(t)dt), \ldots, (\tau+n\delta, \frac{1}{\delta}\int_{\tau+(n-1)\delta}^{\tau+n\delta} S(t)dt)\}$. Els resultats que cal calcular són $r_1 = \frac{1}{t_M-t_m}\int_{t_m}^{t_M} S(t)dt$ i $r_2 = \frac{1}{t_M-t_m}\int_{t_m}^{t_M} S'(t)$.

  Si suposem $\tau = t_m$ i $\tau + n\delta = t_M$ aleshores $\int_{t_m}^{t_M} S'(t) = \int_\tau^{\tau+\delta} S'(t)dt +$

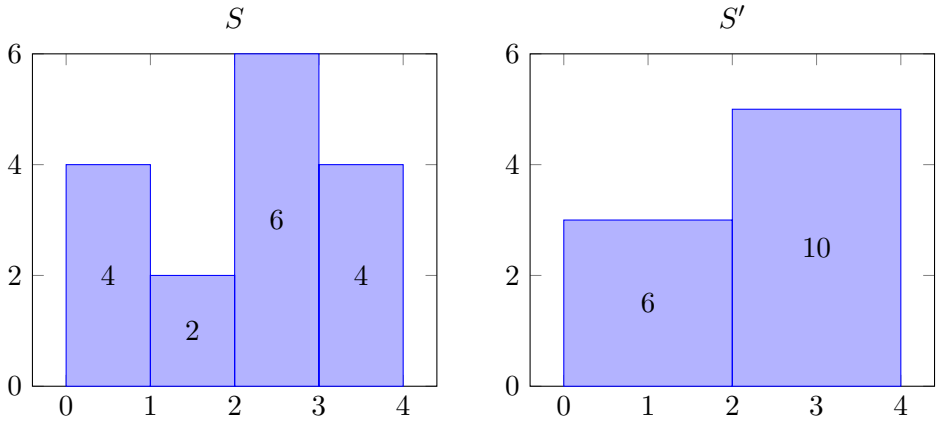


Figura 9.3.: Sèrie temporal amb àrea sota la corba i sèrie temporal resultant de la multiresolució amb agregació mitjana de la funció

$\int_{\tau+\delta}^{\tau+2\delta} S'(t)dt + \cdots + \int_{\tau+(n-1)\delta}^{\tau+n\delta} S'(t)dt$.

Si resolem per exemple el primer interval de consolidació: $\int_{\tau}^{\tau+\delta} S'(t)dt = \delta V(m'_0) = \delta\frac{\int_{\tau}^{\tau+\delta} S(t)dt}{\delta}$ on m'_0 és la mesura corresponent a la consolidació en el primer interval. Així doncs $\int_{\tau}^{\tau+\delta} S'(t)dt = \int_{\tau}^{\tau+\delta} S(t)dt$, de fet és la propietat que resumeix la mitjana de la funció, i per tant podem estendre-ho a $\int_{t_m}^{t_M} S'(t) = \int_{t_m}^{t_M} S(t)$. Podem concloure, doncs, que $r_1 = r_2$.

A la figura 9.3 es mostra un exemple amb valors concrets on $S = \{(1,4),(2,2),(3,6),(4,4)\}$ i l'esquema de multiresolució és $E = \{(\delta = 2, f = \text{mitjana}^{\text{ZOHE}}, \tau = 0, \kappa = \infty)\}$, segons la mitjana$^{\text{ZOHE}}$ de la la definició 5.38. La sèrie resultant de la consolidació de la multiresolució és $S' = \{(2,3),(4,5)\}$. A la figura, les sèries temporals es representen amb representació ZOHE, la superfície pintada en blau correspon a l'àrea de sota la corba de la sèrie temporal i els nombres interiors indiquen el valor d'aquesta àrea; en la sèrie temporal S les àrees es corresponen amb els valors de la sèrie temporal perquè els intervals són d'una unitat de temps. Així doncs es pot observar a S' com aquest esquema de multiresolució conserva el comptatge total en la nova resolució, és a dir $6 = 4 + 2$ en el primer interval consolidat i $10 = 6 + 4$ en el segon, i en una consulta del comptatge total per a tot l'interval $[0,4]$ s'obté 16 tant en S com a S'. Aquesta és la idea bàsica de conservació d'informació en els comptadors.

La multiresolució, però, no pot conservar la resolució del comptatge, així en l'exemple un cop s'ha consolidat $6 = 4 + 2$ no es pot tornar a obtenir 4 i 2 llevat que es pogués reconstruir el senyal. Com s'ha exposat a l'apartat 9.3.3, en consultes en què l'interval no es correspongui amb les resolucions emmagatzemades, els totals no seran els correctes que s'obtindrien de calcular amb les dades originals. De fet,

és el mateix problema que hem exposat que a partir d'un comptador no es poden deduir exactament energies que no s'han mesurat.

9.3.5. Equivalències en l'agregació d'atributs

Hi ha casos en què és el mateix aplicar una funció d'agregació d'atributs que aplicar-ne una altra. En aquest apartat particularment estudiem el cas de la mitjana$^{\text{PD}}$ (v. def. 5.36) i el de la mitjana$^{\text{ZOHE}}$ (v. def. 5.38) per a sèries temporals regulars.

Sigui $S = \{m_0, \ldots, m_k\}$ una sèrie temporal regular de període p (v. def. 4.59) i sigui $[s,t]$ un interval de temps on $s = T(\min(S)) - p = t_0 - p$ i $t = T(\max(S)) = t_k$. Demostrem que $\text{mitjana}^{\text{ZOHE}}(S,[s,t]) = \text{mitjana}^{\text{PD}}(S,[s,t])$ pel que fa al valor resultant calculat negligint el temps resultant. És a dir en realitat demostrem $V(\text{mitjana}^{\text{ZOHE}}(S,[s,t])) = V(\text{mitjana}^{\text{PD}}(S,[s,t]))$ però no escrivim la projecció de l'atribut de valor, $V()$, per a no complicar la notació.

La mitjana aritmètica de la sèrie temporal és $\text{mitjana}^{\text{PD}}(S,[s,t]) = \frac{v_0+\cdots+v_k}{|S|}$. La mitjana amb representació ZOHE és $\text{mitjana}^{\text{ZOHE}}(S,[s,t]) = \frac{1}{t-s}(v_0(t_0-s) + v_1(t_1 - t_0) + \cdots + v_k(t_k - t_{k-1}))$.

Per ser regular, $t_1 - t_0 = \cdots = t_k - t_{k-1} = p$. A més $s = t_0 - p$ i per tant $t_0 - s = p$. També per ser regular, $t_k - t_0 = t_k + (-t_{k-1} + t_{k-1}) + \cdots + (-t_1 + t_1) - t_0 = (|S|-1)p$ i per tant $t - s = t_k - (t_0 - p) = (|S| - 1)d + p = |S|p$.

Reescrivint, $\text{mitjana}^{\text{ZOHE}}(S,[s,t]) = \frac{1}{|S|p}(v_0 p + v_1 p + \cdots + v_k p) = \frac{1}{|S|}(v_0 + v_1 + \cdots + v_k) = \text{mitjana}^{\text{PD}}(S,[s,t])$.

Així doncs, en una sèrie temporal regular es pot aplicar la mitjana$^{\text{PD}}$ com a equivalent a la mitjana$^{\text{ZOHE}}$; de fet la figura 9.3 n'és un exemple. Un àmbit d'aplicació d'aquestes equivalències pot ser el de simplificar els càlculs en les consultes als SGSTM, en els discs dels quals les subsèries temporals normalment s'emmagatzemen regulars.

Part IV.

Experimentació

10. Introducció a les implementacions

En aquest capítol implementem SGST i SGSTM segons els models definits.

Idealment, en un SGBD l'usuari executa una consulta i l'optimitzador s'encarrega d'executar les operacions físiques més eficients, en aquest sentit el model relacional permet trobar expressions equivalents donada una determinada consulta. Això no obstant, i com s'ha detallat en la secció 2.2, una implementació de SGBD no pot abastar i ser eficient en tots els àmbits i contextos. Així doncs, no és tan senzill mantenir totalment la independència entre l'usuari i la implementació ja que cal que aquest decideixi un SGBD adequat per a cada context i fins i tot que declari com resoldre algunes operacions. Ara bé, és possible mantenir la independència del nivell lògic respecte a les implementacions, i és en aquest sentit que a continuació avaluem diferents implementacions per als models definits de SGST i de SGSTM, on cada implementació està pensada per a un context determinat.

Anomenem de forma diferent cada implementació que dissenyem per tal de distingir-les clarament. Són les següents:

- Pytsms i RoundRobinson: Implementacions a alt nivell per a observar el funcionament a nivell acadèmic, amb llenguatge Python. Aquesta és la nostra implementació de referència per als models de SGST i SGSTM, la qual a més usem per als experiments amb dades.
- RoundRobindoop: Implementació específica per a la resolució en diferit i amb computació paral·lela de la multiresolució, amb model de programació MapReduce i en el sistema de computació distribuïda Hadoop.
- Reltsms: Implementació a alt nivell en un SGBDR, amb llenguatge Tutorial D i en el sistema Rel.

Finalment, en un exemple complet s'experimenta amb les implementacions amb dades reals. Les implementacions són de programari lliure i els codis font es poden trobar a [83].

10.1. Particularitats de les implementacions

Els models d'implementacions pertanyen al nivell físic dels SGBD i són una realització d'un model lògic, en el nostre cas dels models lògics de SGST i de SGSTM. Per

a les implementacions se sol definir el nivell d'usuari, que és el llenguatge que serà visible per als usuaris. Les implementacions que realitzem volen ser molt properes al model lògic i per tant el nivell d'usuari que se'n deriva és molt similar. Per a ser un llenguatge d'usuari complet es requereixen facilitats i capacitats de llenguatge de programació –bucles, condicionals, declarar variables, etc.– per a la qual cosa ens basem en els recursos particulars de cada implementació: Python, Tutorial D, etc. L'objectiu principal, però, d'aquestes implementacions és acadèmic i per tant no considerem prioritari el llenguatge d'usuari.

En la implementació s'afegeixen alguns operadors que en el model estructural no havíem definit explícitament perquè ja són propis de l'àlgebra de conjunts. Així per exemple, alguns d'aquests operadors que s'han d'implementar són els relacionats amb la notació de creació de conjunts, els quals en els SGBD s'inclouen en el que es coneix com a llenguatge de definició de dades (DDL, de l'angl. *data definition language*), o els relacionats amb la manipulació de les dades amb assignació, inserció, modificació o esborrament, els quals en els SGBD es coneixen com a llenguatge de manipulació de dades (DML, de l'angl. *data manipulation language*). En el cas dels SGSTM sí que hem definit en el model algunes operacions de DDL i DML per a l'esquema multiresolució, ja que aquest requereix ser manipulat coherentment.

10.2. Qüestions de format

En els capítols següents s'utilitzen llistats per a mostrar exemples de funcionament i els resultats de les operacions efectuades. El format dels llistats depèn del tipus d'informació que mostren. Així, bàsicament, els exemples de codi tenen fons gris, com el llistat 10.1, i els exemples de fitxers o sortides per pantalla tenen fons groc, com el llistat 10.2.

Llistat 10.1: Exemple de codi

```
>>> from pytsms import TimeSeries, Measure as m

>>> s = TimeSeries([m(1,6),m(5,2),m(8,5),m(10,0),m(14,1),m(19,6),m(22,11),m(26,6),m(29,0)])
```

Llistat 10.2: Exemple de fitxer o de sortida per pantalla

```
10/maximum_zohe-10      10  0.0
10/maximum_zohe-10      14  1.0
```

11. Implementació de referència

La implementació de referència dels models de SGST i de SGSTM es realitza amb Python [106]. L'objectiu d'aquesta implementació de referència és mantenir la fidelitat al model per tal de poder experimentar-hi amb tota la potència matemàtica. Implementem els dos models de SGST i SGSTM com a dues biblioteques diferents: *Pytsms* i *RoundRobinson* respectivament. La RoundRobinson, però, està fortament relacionada amb la Pytsms de la mateixa manera que hem definit els SGSTM en base als SGST.

En la implementació amb Python utilitzem el paradigma de programació d'orientació a objectes. Aquest paradigma permet definir objectes que tenen atributs per a estructurar les dades i mètodes associats per a manipular-les, cosa que ens permet relacionar de manera senzilla la implementació amb els conceptes d'estructura i d'operacions del model. Per a mostrar les relacions entre els diversos objectes utilitzem diagrames de classe de Llenguatge unificat de modelització (UML, de l'angl. *Unified Modeling Language*). Aquests diagrames mostren els objectes en caixes però, per tal que no esdevinguin massa grans, només hi assenyalem el nom de l'objecte, és a dir sense els atributs ni els mètodes.

Implementem la part essencial dels models, és a dir l'àlgebra definida en els models lògics. No implementem complements habituals dels SGBD que podrien ser necessaris en entorns d'explotació, com per exemple gestió d'usuaris i permisos, còpies de seguretat, llenguatges estàndards de consulta, etc. Tampoc es tenen en compte paràmetres de rendiment; per exemple en el cas de dues subsèries resolució que tenen el mateix pas de consolidació i les funcions d'agregació d'atributs tenen la mateixa funció de representació –per exemple mitjana$^{\text{ZOHE}}$ i màxim$^{\text{ZOHE}}$–, aquestes podrien compartir la mateixa operació de selecció temporal d'interval ZOHE.

11.1. Pytsms

La biblioteca Pytsms implementa un SGST de referència. Així doncs, seguint el model, els objectes principals són les mesures, les sèries temporals i les representacions de les sèries temporals. Tots tres s'implementen respectivament com a classes Measure, TimeSeries i Representation.

La figura 11.1 mostra amb un diagrama UML la relació entre aquests tres objectes principals. Així, per una banda, una TimeSeries té una relació d'agregació amb

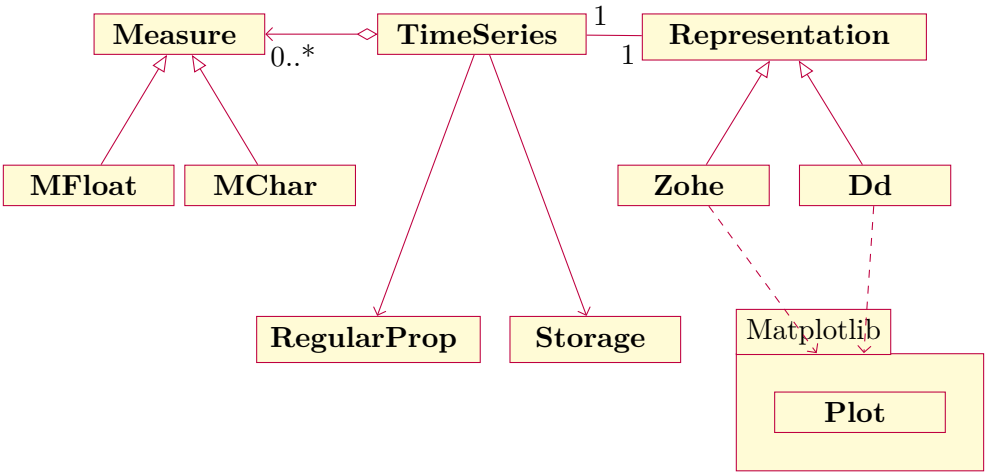


Figura 11.1.: Diagrama UML de Pytsms

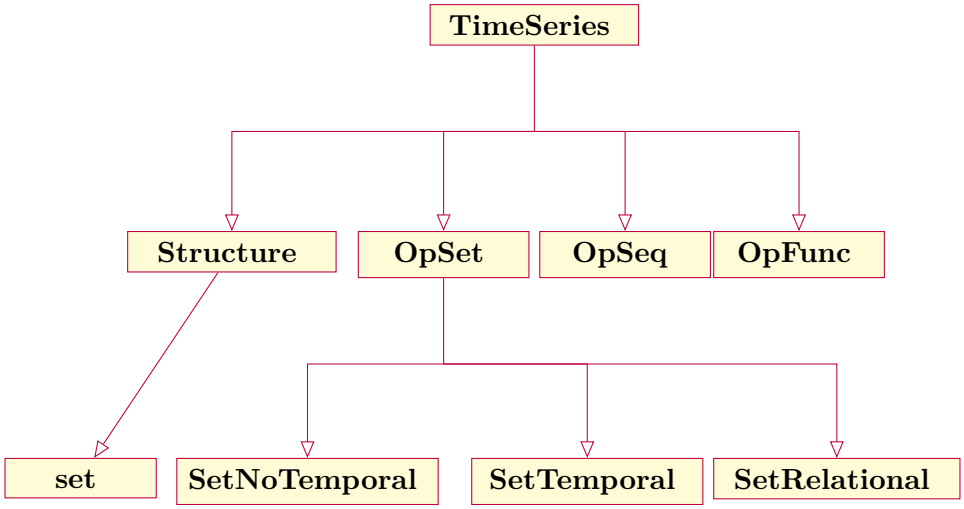


Figura 11.2.: Diagrama UML de la realització de sèries temporals a Pytsms

les Measure, és a dir que una sèrie temporal conté cap, una o més d'una mesura. Per altra banda, les sèries temporals i les representacions són ortogonals i això s'implementa mitjançant una relació d'associació bidireccional entre una TimeSeries i una Representation, és a dir que una instància de sèrie temporal té associada una representació i una instància de representació coneix la sèrie temporal que representa.

Una TimeSeries és un objecte amb una gran quantitat de mètodes. Com a conseqüència, la implementació de funcionalitats essencials s'ha dividit en diversos objectes, els quals es mostren a la figura 11.2. Els mètodes que implementen el model estructural i el model d'operacions bàsiques s'han agrupat en objectes segons la seva funcionalitat. Així hi ha l'objecte Structure que implementa el model estructural de les sèries temporals, l'OpSet pel model d'operacions de conjunts, l'OpSeq pel model d'operacions de seqüències i l'OpFunc pel model d'operacions de funció temporal. Aleshores l'objecte TimeSeries multihereta les funcionalitats d'aquests quatre objectes, cosa que s'implementa a Python com a *Mixin* [106, §8.3.6, §20.17]. L'objecte OpSet també té una gran quantitat de mètodes i, de la mateixa manera, hereta la funcionalitat de tres objectes: el SetNoTemporal per a les operacions basades en l'orde parcial de les sèries temporals, el SetTemporal basat en l'ordre temporal i el SetRelational per a les operacions específiques de l'àlgebra relacional. Pel que fa a l'Structure hereta funcionalitats dels objectes **set**, que són un tipus predefinit a Python [106, §5.7].

Les funcionalitats complementàries de les TimeSeries s'han implementat amb relacions d'associació unidireccionals, les quals es mostren a la figura 11.1. Així, hi ha dues funcionalitats complementàries: RegularProp és un objecte que agrupa les operacions relacionades amb la regularitat de les sèries temporals i Storage agrupa les operacions d'emmagatzematge i de recuperació en fitxers. En aquests casos, l'associació unidireccional indica que les TimeSeries són objectes que admeten les operacions complementàries i s'implementa a Python seguint el patró de disseny *Visitor* [90, 140]. Amb aquest patró les TimeSeries esdevenen *Visitable*, és a dir accepten objectes *Visitor* que aporten funcionalitats extres. Així doncs, els objectes RegularProperties i Storage són *Visitor* i les TimeSeries tenen un mètode accept que permet acceptar-los.

La figura 11.1 mostra exemples d'especialitzacions de les mesures i de les representacions. Pel que fa a les Measure, poden tenir especialitzacions segons els tipus dels atributs de temps i de valor. Amb aquesta relació implementem la propietat homogènia de les sèries temporals i la definició de mesura indefinida i de valor indefinit, és a dir que totes les mesures que conté una sèrie temporal són del mateix tipus i cada tipus de mesura té uns valors de l'atribut temps que la defineixen indefinida i uns valors de l'atribut valor que la defineixen de valor indefinit. Així, per defecte, una Measure defineix els reals $-\infty$ i $+\infty$ per a les mesures indefinida negativa i positiva respectivament, i defineix el valor None de Python per a la mesura de valor indefinit. Aleshores, mitjançant especialitzacions es poden definir altres tipus de

mesures; per exemple la MFloat que defineix el real ∞ com a valor indefinit o bé la MChar que defineix mesures de tipus caràcter.

Pel que fa a les representacions, cada representació en concret és una especialització de Representation. Per exemple Zohe i Dd implementen la funció de representació ZOHE i la DD respectivament. Bàsicament, cada representació particular ha de definir l'operació que calcula l'interval temporal i l'operació que permet trobar-ne el graf. També cadascuna implementa una operació que dibuixi correctament el gràfic de la sèrie temporal segons la representació; en les representacions definides s'usa la biblioteca *Matplotlib* [126] per a fer els gràfics.

11.2. RoundRobinson

La biblioteca RoundRobinson implementa un SGSTM de referència. Així doncs, seguint el model, els objectes principals són les sèries temporals multiresolució, les subsèries resolució, els discs, els buffers i les funcions d'agregació d'atributs. Respectivament s'implementen com a classes MultiresolutionSeries, Resolution, Disc, Buffer i Function –les funcions d'agregació d'atributs són realitzades per Function de Python.

La figura 11.3 mostra amb un diagrama UML la relació entre aquests cinc objectes principals. Així, una MultiresolutionSeries té una relació d'agregació amb les Resolution, és a dir que una sèrie temporal multiresolució conté subsèries resolucions. Una Resolution té una relació de composició amb un Buffer i una altra amb un Disc, és a dir que cada subsèrie resolució està formada exactament per un buffer i un disc. Cada Buffer té una relació d'associació amb una TimeSeries, és a dir amb la sèrie temporal del buffer; de manera similar per la sèrie temporal del disc cada Disc s'associa a una TimeSeries. A més, cada Buffer també té una relació d'associació amb una Function que ha de tenir dos paràmetres: la sèrie temporal (s) i l'interval de consolidació (i). Noteu que en el model hem definit una funció d'agregació d'atributs sobre una sèrie temporal, un instant de consolidació τ i un pas de consolidació δ, mentre que les implementem sobre una sèrie temporal i directament sobre l'interval de consolidació $i = [\tau, \tau + \delta]$.

Les MultiresolutionSeries tenen funcionalitats complementàries que s'han implementat amb relacions d'associació unidireccionals. Així, hi ha dues funcionalitats complementàries: Plot per a les operacions relacionades amb la visualització gràfica i Storage per a operacions d'emmagatzematge i de recuperació en fitxers. Aquests dos objectes també s'han implementat amb patró de disseny *Visitor* i les MultiresolutionSeries són *Visitable* de la mateixa manera que en el cas de les funcionalitats complementàries de Pytsms.

Les MultiresolutionSeries com a conjunts formats per Resolution s'han implementat heretant funcionalitats dels **set** de Python. Així, per a definir l'esquema multiresolu-

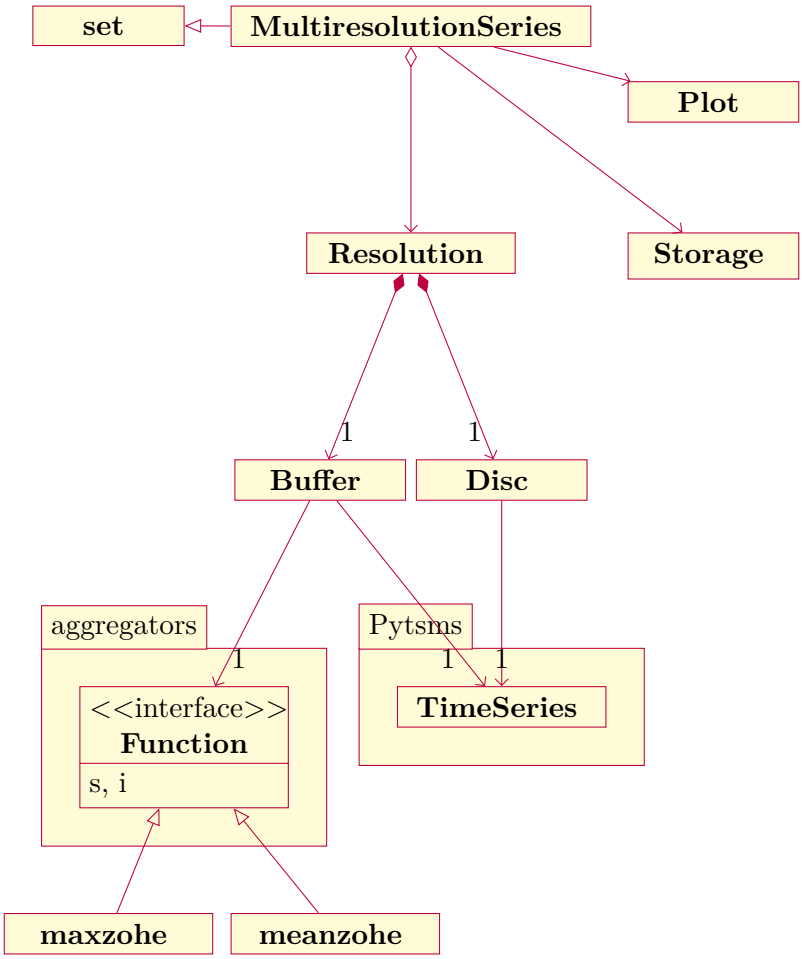


Figura 11.3.: Diagrama UML de RoundRobinson

ció hi ha un mètode addResolution que permet afegir subsèries resolució. Cadascuna es configura amb quatre paràmetres: delta, k, f i tau, els quals creen una Resolution amb el Buffer i el Disc corresponent.

Les MultiresolutionSeries tenen mètodes que operen amb totes les Resolution contingudes. Els més importants són el mètode add per a afegir noves mesures, el consolidable per a determinar si alguna subsèrie és consolidable i el consolidate que consolida totes les consolidables.

Per a les consultes, hi ha dos mètodes: discSeries i total. El discSeries té com a paràmetres delta i f i retorna la TimeSeries del Disc corresponent. El total retorna la concatenació de tots els discSeries possibles ordenats per delta; com que no hi pot haver delta repetits, el mètode total té un paràmetre f que permet seleccionar només aquells discSeries que tenen unes determinades funcions d'agregació d'atributs.

En un mòdul aggregators hi ha predefinides algunes funcions d'agregació d'atributs com per exemple la max_zohe o la mean_zohe que respectivament agreguen el màxim i la mitjana seguint la representació ZOHE. Això no obstant, els usuaris en poden definir de pròpies ja que són Function de Python amb els paràmetres s una TimeSeries i i una parella d'instants de temps com ara [ta,tb] i que retornen una Measure. Per exemple la funció mean_zohe retorna una mesura amb tb com a temps i com a valor l'àrea mitjana a partir la selecció temporal ZOHE de la sèrie temporal en l'interval de temps donat.

11.3. Exemples d'ús

Amb les biblioteques Pytsms i RoundRobinson podem treballar amb les sèries temporals i les sèries temporals multiresolució de manera molt semblant als models algebraics de SGST i de SGSTM.

El llistat 11.1 és un exemple d'ús de Pytsms on definim dues sèries temporals s1 i s2 i hi apliquem diverses operacions: unió, unió temporal, concatenació, interval tancat, interval temporal ZOHE, selecció temporal ZOHE, comprovació de propietats de regularitat i consultes de gràfics. Noteu que Measure s'abrevia amb m.

Llistat 11.1: Exemple d'operacions amb Pytsms

```
#Importació dels objectes necessaris
>>> from pytsms import TimeSeries, Measure as m
>>> from pytsms.representation import Zohe
>>> from pytsms.properties import isRegular

#Definició de les dues sèries temporals d'exemple
>>> s1 = TimeSeries([ m(1,1), m(3,1), m(4,0), m(5,1) ])
>>> s2 = TimeSeries([ m(2,2), m(3,2), m(4,0), m(6,2) ])

#Manipulacions de les dues sèries temporals
```

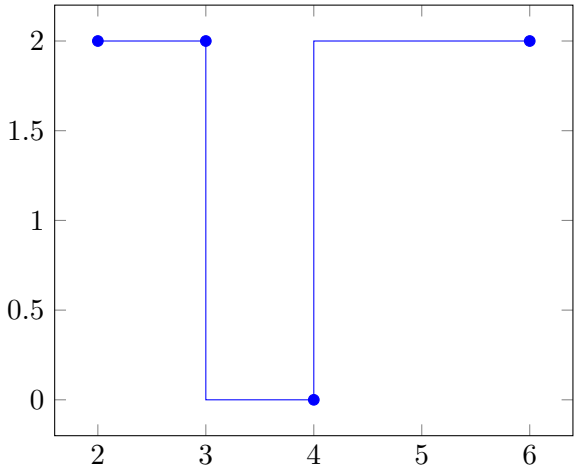


Figura 11.4.: Gràfic de la sèrie temporal d'exemple s_2 amb representació ZOHE

```
# s_1 ∪ s_2
>>> s1.union(s2)
TimeSeries([m(1,1), m(2,2), m(3,1), m(4,0), m(5,1), m(6,2)])
# s_1 ∪^t s_2
>>> s1.union_temporal(s2)
TimeSeries([m(1,1), m(2,2), m(4,0), m(5,1), m(6,2)])
# s_1||s_2
s1.concatenate(s2)
TimeSeries([m(1,1), m(3,1), m(4,0), m(5,1), m(6,2)])
# s_2[2, 5]
>>> s2.interval_closed(2,5)
TimeSeries([m(2,2), m(3,2), m(4,0)])
# s_2[2, 5]^ZOHE
>>> s2.interval_temporal(2,5,Zohe)
TimeSeries([m(3,2), m(4,0), m(5,2)])

#Comprovació de la regularitat
# s_2 no és regular
>>> s2.accept(isRegular())
False
# regularitzem s_2 amb la selecció temporal s_2[0, 2, 4]^ZOHE
>>> r2 = s2.selection_temporal([0,2,4],Zohe)
>>> r2
TimeSeries([m(0,2), m(2,2), m(4,0)])
>>> r2.accept(isRegular())
True

Gràfic de la sèrie temporal s_2 amb representació ZOHE
>>> s2.plot(rpr=Zohe)
[v. figura 11.4]
```

Al llistat 11.1 es mostra l'ús de funcionalitats complementàries de Pytsms, el qual s'ha dissenyat extensible per a poder incorporar noves operacions, per a la regulari-

tat. En el llistat 11.2 cas es mostra l'ús d'operacions d'emmagatzematge en fitxers, del mòdul storage. Primer s'emmagatzema la s2 del llistat 11.1 en un fitxer amb format de valors separats per comes (CSV, de l'angl. *comma-separated values*), en què cada línia és una parella de temps i valor separats per una coma, i després es recupera del fitxer la sèrie temporal emmagatzemada. Les dades emmagatzemades al fitxer es mostren al llistat 11.3.

Llistat 11.2: Operacions complementàries de Pytsms per a l'emmagatzematge

```
#Importació dels objectes necessaris
>>> from pytsms.storage import SaveCsv, LoadCsv

#Emmagatzematge en format CSV de nom st2.csv
>>> s2.accept(SaveCsv('st2.csv'))
#Recuperació a partir de format CSV
>>> sr = TimeSeries([])
>>> sr.accept(LoadCsv('st2.csv'))
TimeSeries([m(2,2), m(3,2), m(4,0), m(6,2)])
```

Llistat 11.3: Dades del fitxer st2.csv

```
2,2
3,2
4,0
6,2
```

El llistat 11.4 és un exemple d'ús de RoundRobinson on definim una sèrie temporal multiresolució M, hi afegim una sèrie temporal s, i hi apliquem la consolidació fins que no sigui consolidable. Finalment, consultem el resultat amb les dues consultes bàsiques –SèrieDisc i SèrieTotal– i ho dibuixem gràficament. Les dades són les mateixes que a l'exemple 5.1 i també la consolidació resultant, tot i que ara les sèries temporals tenen valors diferents de zero per a poder observar els resultats de les funcions d'agregació d'atributs.

Llistat 11.4: Exemple d'operacions amb RoundRobinson

```
#Importació dels objectes necessaris
>>> from pytsms import TimeSeries, Measure as m
>>> from roundrobinson import MultiresolutionSeries
>>> from roundrobinson.aggregators import mean_zohe,maximum_zohe
>>> from roundrobinson.plot import Plot
>>> from pytsms.representation import Zohe

#Definició de la sèrie temporal d'exemple
>>> s = TimeSeries([m(1,6),m(5,2),m(8,5),m(10,0),m(14,1),m(19,6),m(22,11),m(26,6),m(29,0)])

#Definició de la sèrie temporal multiresolució
>>> M = MultiresolutionSeries()
#Definició de l'esquema multiresolució
>>> M.addResolution(delta=5,k=4,f=mean_zohe,tau=0)
>>> M.addResolution(delta=10,k=3,f=maximum_zohe,tau=0)
```

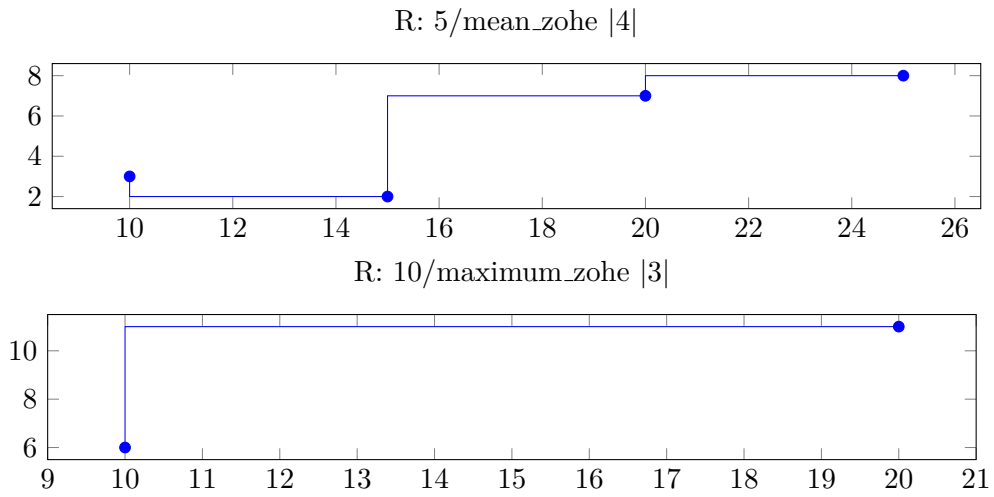


Figura 11.5.: Base de dades multiresolució M

```
#Addició de totes les mesures de la sèrie temporal
>>> for m in s: M.add(m)
#M ja és consolidable
>>> M.consolidable()
True
#Consolidació fins que no sigui consolidable
>>> while M.consolidable(): M.consolidate()

#Consulta SèrieDisc(M, 5, mitjana^ZOHE)
>>> M.discSeries(5,mean_zohe)
TimeSeries([m(10,3), m(15,2), m(20,7), m(25,8)])
#Consulta SèrieDisc(M, 10, màxim^ZOHE)
>>> M.discSeries(10,maximum_zohe)
TimeSeries([m(10,6), m(20,11)])
#Consulta SèrieTotal(M)
>>> M.total()
TimeSeries([m(10,3), m(15,2), m(20,7), m(25,8)])
#Gràfic multiresolució
>>> M.accept(Plot())
[v. figura 11.5]
#Gràfic Pytsms de la sèrie temporal total amb representació ZOHE
>>> M.total().plot(rpr=Zohe)
[v. figura 11.6]
```

Al llistat 11.5 es mostra l'ús de funcionalitats complementàries de RoundRobinson, el qual s'ha dissenyat extensible per a poder incorporar noves operacions, per als gràfics. En el llistat 11.5 es mostra l'ús d'operacions d'emmagatzematge en fitxers, del mòdul storage. Primer s'emmagatzema la M del llistat 11.5 en un fitxer amb format Pickle, que és un format d'emmagatzematge de Python [106], i després es recupera del fitxer la sèrie temporal multiresolució emmagatzemada. Segon s'em-

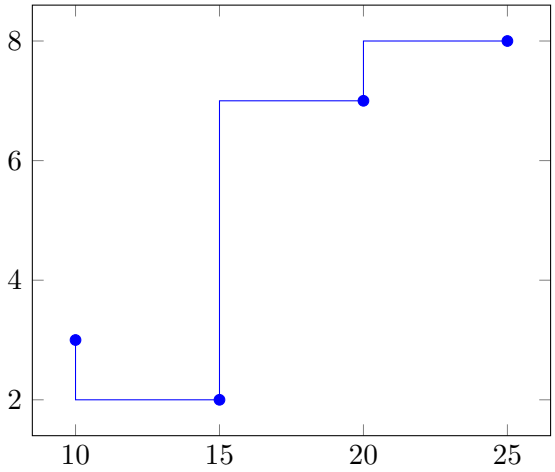


Figura 11.6.: SèrieTotal(M) amb representació ZOHE

magatzema la M del llistat 11.5 en un fitxer amb format CSV i després es recupera del fitxer la sèrie temporal multiresolució emmagatzemada. Les dades emmagatzemades al fitxer CSV es mostren al llistat 11.6.

Llistat 11.5: Operacions complementàries de RoundRobinson per a l'emmagatzematge

```
#Importació dels objectes necessaris
>>> from roundrobinson.storage import SavePickle, LoadPickle, SaveCsv, LoadCsv

#Emmagatzematge en format Pickle de nom mrd.pickle
>>> M.accept(SavePickle('mrd.pickle'))
#Recuperació a partir de format Pickle
>>> Mr = MultiresolutionSeries([])
>>> Mr = Mr.accept(LoadPickle('mrd.pickle'))
>>> Mr == M
True

#Emmagatzematge en format CSV de nom mrd.csv
>>> M.accept(SaveCsv('mrd.csv'))
#Recuperació a partir de format CSV
>>> Mr = M.accept(LoadCsv('mrd.csv'))
```

Llistat 11.6: Dades del fitxer mrd.csv

```
5 mean_zohe 25 8
5 mean_zohe 10 3
5 mean_zohe 20 7
5 mean_zohe 15 2
10 maximum_zohe 10 6
10 maximum_zohe 20 11
```

Cal notar que en el cas de l'emmagatzematge en CSV no es tenen en compte els buffers, per tant la recuperació no és estrictament el mateix que la sèrie temporal multiresolució original.

12. Implementació amb paral·lelisme

En el capítol 7 hem definit la funció de multiresolució com una funció sobre una sèrie temporal. Aquesta funció té bàsicament dues parts: un mapa sobre uns paràmetres de multiresolució i un plec sobre un esquema de multiresolució. Ara implementem aquestes funcions mitjançant computació paral·lela.

Una tècnica de computació paral·lela és MapReduce [34], la qual s'adequa bé al problema ja que es basa en aplicar operacions de mapa (*map*) i posteriorment plegar-les (*reduce*). Un sistema que es basa exclusivament en aquesta tècnica és Hadoop [124].

A continuació, en primer lloc, estudiem Hadoop i la tècnica MapReduce. En segon lloc, implementem usant Hadoop un SGSTM anomenat *RoundRobindoop*. Aquest és un SGSTM específic amb l'objectiu de mostrar una implementació que resolgui la multiresolució d'una sèrie temporal en temps diferit (*offline*) i computant paral·lelament.

12.1. Hadoop i MapReduce

Apache Hadoop, o simplement Hadoop, [124] és un sistema de computació distribuïda que permet processar grans volums de dades amb diferents computadors en paral·lel. El sistema inclou la gestió de l'emmagatzematge per a distribuir les dades als diferents computadors, la qual cosa s'anomena *Hadoop Distributed File System* (HDFS); la gestió dels diferents processos en els diversos computadors; i el model de programació paral·lela, el qual és MapReduce.

MapReduce [34, 76] és un model de programació per processar algoritmes en paral·lel. Es basa en resoldre els algoritmes en dues etapes: primer en una etapa de *maps* i segon en una etapa de *reduces*. Aquestes dues etapes són l'algoritme bàsic i per això s'anomena MapReduce, tot i que hi ha variacions que afegeixen més etapes. Els noms de *map* i *reduce* també s'usen per a les operacions d'alt ordre, com les mapa (*map*) i plec (*fold* o també *reduce*), que hem definit en el model dels SGST, però Lämmel [76] compara les de MapReduce amb les d'alt ordre i conclou que no són exactament el mateix; aquí distingirem els conceptes usant els noms en anglès map i reduce per a MapReduce.

A la figura 12.1 es mostra l'esquema de funcionament de MapReduce, que és el següent:

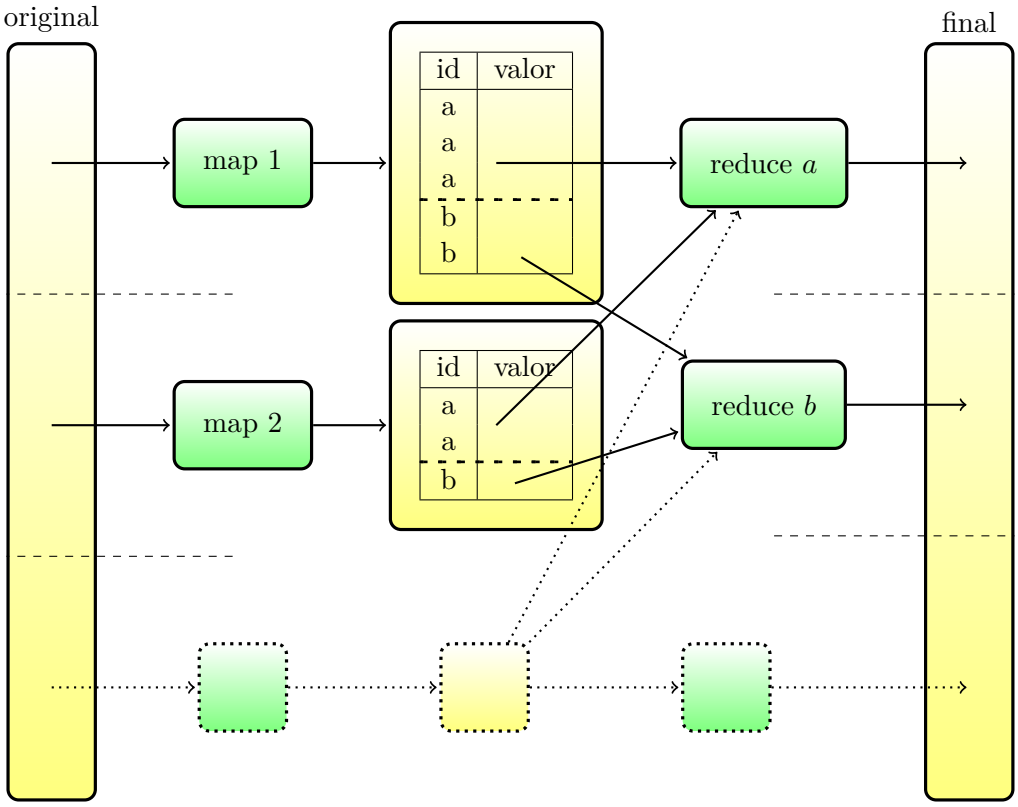


Figura 12.1.: Esquema de funcionament de MapReduce

1. Hi ha unes dades originals que es poden partir en trossos. Hadoop està orientat a fitxers, mitjançant HDFS, i per tant cada tros de dades és cadascun dels fitxers que es volen processar o bé conjunts de línies d'un fitxer.
2. Cada tros de les dades es processa mitjançant una operació map. Cada map es pot computar en paral·lel i distribuït.
3. Cada operació map ha de retornar un nou conjunt de dades formats per parelles d'identificador i valor. Aquests conjunts de dades s'ordenen per identificador.
4. Cada conjunt de dades amb el mateix identificador es processa mitjançant una operació reduce. Cada reduce es pot computar en paral·lel i distribuït.
5. Cada reduce ha de retornar un tros del resultat final. És a dir, que unint les dades que retornen els reduce s'obtenen les dades finals. En l'orientació a fitxers de Hadoop, el resultat final és un fitxer, o bé un tros d'un fitxer, per a cada reduce.

Per a resoldre un algoritme amb MapReduce, cal definir l'operació de map i l'operació de reduce. El map ha de calcular un filtre sobre les dades, sobretot establir grups de dades, i el reduce ha de calcular agregacions o resums per a cada grup. MapReduce té una gran similitud amb l'operació *summarize* dels SGBDR [23, cap. 7], però separada convenientment en les dues etapes. A més, MapReduce imposa les següents restriccions: les dades s'han de poder partir i l'algoritme s'ha de poder expressar separat en les dues operacions de map i de reduce. Dean i Ghemawat [34] mostren exemples d'algoritmes que es poden expressar amb MapReduce.

Un cop s'ha modelat un algoritme amb MapReduce, aleshores Hadoop ja és capaç d'executar els maps i els reduces en paral·lel i distribuïts. A més, Hadoop també gestiona el compromís dels recursos entre el temps de distribuir les dades, la quantitat de processos en paral·lel que s'han de crear i el temps afegit que suposa cada procés nou.

12.2. RoundRobindoop

RoundRobindoop implementa un SGSTM específic que resol la funció de multiresolució amb el model de programació MapReduce. Per a construir RoundRobindoop, primer dissenyem l'algoritme de MapReduce que s'adequa a la multiresolució. Segon, proposem dues maneres per a executar RoundRobindoop: amb Hadoop i la shell del sistema.

Cardinal	Conjunt
i	nombre de maps
j	nombre de reduces
k	sèrie temporal (S)
l	esquema multiresolució (E)
n	dades d'entremig (X)
o	dades finals (Y)

Taula 12.1.: Notació dels conjunts i els cardinals per a definir RoundRobindoop

12.2.1. Multiresolució amb MapReduce

L'algoritme que implementem amb MapReduce és el de la funció de multiresolució. Aquesta funció principalment té dues parts segons les definicions realitzades: el mapa de multiresolució (MapMu, v. def. 7.1) i el plec de multiresolució (PlecMu, v. def. 7.2). Com ja s'ha dit, aquestes definicions no es poden correspondre exactament amb les operacions de map i de reduce. Així doncs, dissenyem les operacions de map i de reduce que tenen el mateix efecte que calcular les funcions de multiresolució, en què el resultat final no és la sèrie temporal total sinó el resultat de totes les funcions MapMu, és a dir el resultat final són les sèries temporals dels discs del model de SGSTM la concatenació dels quals resulta en el PlecMu.

Sigui S una sèrie temporal, i $E = \{(\delta_1, f_1, \tau_1, \kappa_1), \ldots, (\delta_l, f_l, \tau_l, \kappa_l)\}$ un esquema de multiresolució, definim l'algoritme MapReduce que calcula mapreduce(S, E) $= \{(\delta_1, f_1, \text{MapMu}(S, \delta_1, f_1, \tau_1, \kappa_1)), \ldots, (\delta_l, f_l, \text{MapMu}(S, \delta_l, f_l, \tau_l, \kappa_l))\}$. És a dir, calcula tots els MapMu possibles i els identifica amb el pas de consolidació δ i la funció d'agregació d'atributs f, els quals identifiquen les subsèries resolució assumint que no n'hi ha de repetits.

L'esquema de funcionament de RoundRobindoop és el de la figura 12.2, el qual és la implementació particular de l'esquema de la figura 12.1 per a la multiresolució. De forma resumida, RoundRobindoop en l'etapa de map classifica les mesures en funció de a quin disc i temps resultant es consolidaran i en l'etapa de reduce calcula la funció d'agregació per a les mesures que s'hagin de consolidar al mateix disc. A continuació expliquem detalladament com són les dades originals (S), les d'entremig (X) i les finals (Y), i finalment definim l'operació de map i la de reduce. Per a les dades d'entremig, distingim com resulten de l'operació map i com s'agrupen per a l'operació reduce. A la taula 12.1 resumim la notació dels conjunts i els seus cardinals utilitzats en aquesta secció.

Dades originals Sigui i el nombre de processos map i sigui la sèrie temporal $S = \{m_0, \ldots, m_a, \ldots, , m_k\}$ les dades originals. La sèrie temporal es pot partir en subconjunts $S = S_1 \cup S_2 \cup \cdots \cup S_i$ on cada subconjunt és una subsèrie temporal

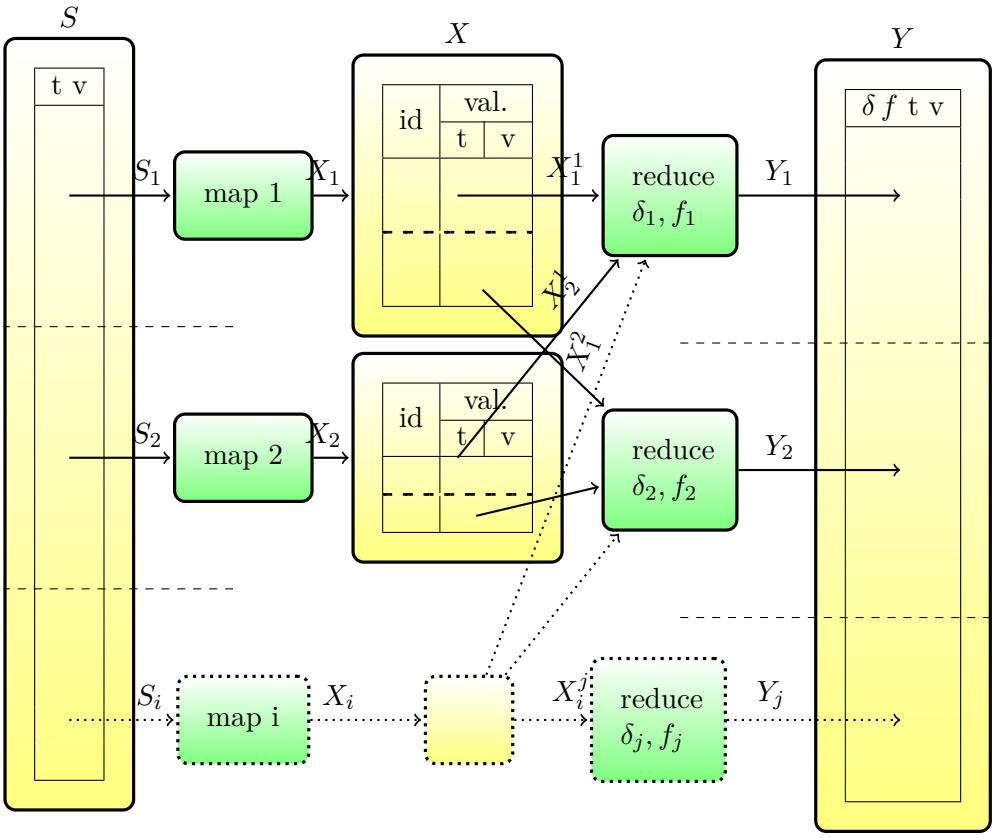


Figura 12.2.: Esquema de funcionament de RoundRobindoop

$S_1 = \{m_0, \ldots, m_a\}, S_2 = \{m_{a+1}, \ldots\}, \ldots, S_i = \{\ldots, m_k\}$ i $a \in \mathbb{N}$, $a < k$. Cada operació map rep una subsèrie temporal de l'original.

Dades d'entremig (map) Les dades d'entremig equivalen a les sèries temporals dels buffers, és a dir les mesures pendents de consolidar per a cada subsèrie resolució. Així doncs, les dades d'entremig són un conjunt de dades pendents de consolidar $X = \{\chi_1, \ldots, \chi_n\}$ on cada dada és un tuple $\chi = (\mathrm{id}, m)$ en què $\mathrm{id} = (\delta, f, \tau + \delta)$ funciona com a identificador i la mesura m funciona com a valor. L'identificador classifica per a cada mesura original m les resolucions on s'hauran de consolidar, identificades pel pas de consolidació δ, per la funció d'agregació d'atributs f i pel temps resultant de consolidació $\tau + \delta$.

Cada operació map calcula un subconjunt de les dades d'entremig, és a dir en total $X = X_1 \cup X_2 \cup \cdots \cup X_i$, on per exemple el primer subconjunt té la forma $X_1 = \{(\mathrm{id}_1, m_0), (\mathrm{id}_2, m_0), \ldots, (\mathrm{id}_j, m_0), (\mathrm{id}_1, m_a), \ldots, (\mathrm{id}_j, m_a)\}$ i per exemple el primer identificador té la forma $\mathrm{id}_1 = (\delta_1, f_1, \tau^* + \delta_1)$ i $\tau^* + \delta_1$ seria l'instant en què s'hauria de consolidar m_0. Així doncs, les dades d'entremig tenen un cardinal $|X| = |E||S|$ és a dir la quantitat de resolucions de l'esquema multiplicat per la quantitat de mesures de la sèrie temporal original.

Dades d'entremig (reduce) Un cop calculades, les dades d'entremig X s'ordenen i s'agrupen per identificadors $(\delta, f, \tau + \delta)$ idèntics. Sigui j el nombre de processos reduce i siguin $X = X^1 \cup X^2 \cup \cdots \cup X^j$ les agrupacions d'identificadors. Cada agrupació de dades està formada per subconjunts que provenen de qualsevol map, $X^1 = X_1^1 \cup X_2^1 \cup \cdots X_i^1$ on per exemple $X_1^1 = \{(\delta_1, f_1, \tau^* + \delta_1, m_0), \ldots, (\delta_1, f_1, \tau^* + \delta_1, m_*)\}$ conté les mesures que s'han processat en el primer map i han estat classificades per a processar-se en el reduce identificat per $(\delta_1, f_1, \tau^* + \delta_1)$. De forma més general, un reduce pot rebre diverses agrupacions d'identificadors idèntics, per exemple tots els que comparteixin δ i f que és el cas que hem dibuixat a la figura 12.2.

Dades finals Les dades finals equivalen a les sèries temporals dels discs, és a dir les mesures consolidades per a cada subsèrie resolució. Així doncs, les dades finals són un conjunt de dades consolidades $Y = \{\upsilon_1, \ldots, \upsilon_o\}$ on cada dada consolidada és un tuple $\upsilon = (\delta, f, t, v)$ identificat per la resolució δ i f a què pertany i amb el valor v consolidat en l'instant temps de la forma $t = \tau + \delta$. Sigui j el nombre de reduce. En la forma més general, cada reduce calcula un subconjunt de les dades finals $Y = Y_1 \cup Y_2 \cup \cdots Y_j$. Per exemple en la figura 12.2, que s'agrupen els identificadors per δ i f, el primer subconjunt té la forma $Y_1 = \{(\delta_1, f_1, \tau^* + \delta_1, v^*), \ldots, (\delta_1, f_1, \tau^{**} + \delta_1, v^{**})\}$ on τ^*, v^*, τ^{**} i v^{**} són els instants i valors que corresponguin.

Atès que en la multiresolució cada disc està afitat per un cardinal màxim, es poden afitar el nombre total de processos reduce. Així siguin κ els cardinals de l'esquema de multiresolució, el cardinal de les dades finals és $|Y| \leq \kappa_1 + \kappa_2 + \cdots + \kappa_l$. En el

cas de la figura 12.2, $|Y| \leq l$ perquè hi ha els reduce agrupats per cada resolució δ i f.

Operació map L'operació map calcula les dades d'entremig a partir de les dades originals i un esquema multiresolució. Treballa en subconjunts de les dades per a així poder-se computar paral·lelament.

Definició 12.1 (Operació map)**.** *Sigui S una subsèrie temporal de les dades originals, E un esquema de multiresolució i X un subconjunt de les dades d'entremig abans de ser ordenades. L'operació* map *l'algoritme MapReduce, notada com a $X = \text{map}(S_1, E)$, resulta en el subconbjunt $X = \bigcup_{\forall m \in S} \text{classifica}(m, E)$.*

La funció classifica *indica per a cada mesura a quins discs s'ha de consolidar:*

$$\begin{aligned}\text{classifica}(m, E) = &\{(\delta, f, \tau + n\delta, T(m), V(m)) | (\delta, f, \tau, \kappa) \in E \\ &\wedge\ \tau + (n-1)\delta < T(m) \leq \tau + n\delta \\ &\wedge\ n \in \mathbb{Z} \wedge T(m) > \tau\}\end{aligned}$$

Cal tenir en compte que $T(m) > \tau$ indica que τ és el temps d'inici de la resolució i per tant no té sentit incloure mesures anteriors.

Hi ha dues restriccions a la funció classifica que hem definit:

- S'assumeix que la f treballa sobre l'interval de la sèrie original $S(\tau + (n-1)\delta, \tau + n\delta]$. En cas que no sigui així per a la f escollida, caldria modificar aquests intervals de classificació. A continuació d'aquest apartat contextualitzem aquest problema de les f.
- No es tenen en compte els cardinals màxims. Si es volen tenir en compte, cal ajustar els τ inicials. A continuació d'aquest apartat descrivim com es poden ajustar els temps d'inici de la consolidació.

Exemple 12.1 (Classificació d'una mesura en les resolucions)**.** Sigui l'esquema de multiresolució $E = \{(\delta_1 = 2, f_1, \tau_1 = 0, \kappa_1 = 4), (\delta_2 = 5, f_2, \tau_2 = 10, \kappa_2 = 3)\}$ i la mesura $m = (25, 1)$, aquesta és classificada per a consolidar-se en les dues resolucions següents: $\text{classifica}(m, E) = \{(2, f_1, t^*, 25, 1), (5, f_2, t^{**}, 25, 1)\}$ on $\tau_1 + (n-1)\delta_1 < 25 \leq \tau_1 + n\delta_1, n \in \mathbb{Z}$ i per tant $t^* = \tau_1 + n\delta_1 = 26$, i de manera semblant es pot calcular $t^{**} = 25$.

Operació reduce Un cop s'han obtingut les dades d'entremig el sistema, com per exemple Hadoop, agrupa els tuples amb el mateix identificador i els processa en un mateix reduce. L'operació reduce calcula les dades finals a partir de les dades d'entremig ordenades, tot i que treballa en subconjunts de les dades per a així poder-se computar paral·lelament.

Definició 12.2 (Operació reduce)**.** *Sigui* $X = \{(\delta, f, \tau, m_0), \ldots, (\delta, f, \tau, m_k)\}$ *un subconjunt de les dades d'entremig ordenades i* Y *un subconjunt de les dades finals. L'operació* reduce *de l'algoritme MapReduce, notada com a* $Y = \text{reduce}(X)$*, resulta en el subconjunt* $Y = \{(\delta, f, \tau, v)\}$ *on* $v = V(f(\{m_0, \ldots, m_k\}, \tau - \delta, \delta))$*.*

De manera senzilla, es pot estendre la definició per a treballar amb diverses agrupacions d'identificadors a les dades d'entremig $X = \{(id_0, m_a), \ldots, (id_0, m_b), (id_1, m_c), \ldots, (id_1, m_d), \ldots\}$ *i retornar el corresponent subconjunt final* $Y = \{(id_0, v_0), (id_1, v_1), \ldots\}$*.*

L'expressió de l'interval $[\tau-\delta, \tau]$ es pot ometre ja que l'operació map ja ha classificat les mesures d'aquest interval. Tot i així l'indiquem per a seguir la forma genèrica $f(S, \tau, \delta)$ de les funcions d'agregació d'atributs. A continuació d'aquest apartat contextualitzem aquest problema de les f.

En conclusió, el map i el reduce no es corresponen exactament amb les definicions de les funcions de MapMu i PlecMu, sinó que el map classifica les mesures de la sèrie temporal segons el buffer que hi correspon i el reduce calcula les mesures consolidades per un disc. És a dir que un map i un reduce equivalen a la funcionalitat de la funció MapMu però el conjunt de tots els maps i reduces equivalen a la funció de PlecMu, sense expressar-ho en forma de sèrie temporal total.

Quant a les f a RoundRobindoop

Al model de SGSTM hem definit de forma genèrica les funcions d'agregació d'atributs com a $m = f(S, \tau, \delta)$ (v. § 5.3). Aquestes funcions principalment realitzen dues operacions: una selecció sobre la sèrie temporal en l'interval de temps en què $[\tau, \tau + \delta]$ i una agregació de les mesures seleccionades.

A RoundRobindoop, l'operació de selecció es duu a terme a l'etapa de map i en canvi l'agregació, a l'etapa de reduce. Així en l'algoritme de MapReduce definit per a RoundRobindoop usem el model de f descrit anteriorment, però per a implementar correctament aquestes funcions cal interpretar-ne el significat per a l'etapa de map i per a la de reduce. És a dir, cada f hauria de tenir dos components: un amb les operacions de selecció per a ser usades en els map i l'altre amb les operacions d'agregació per als reduce.

Així doncs, caldria afegir en els paràmetres de RoundRobindoop l'operació de selecció d'interval per a cada f que s'utilitzi. Però això complica la resolució de l'algoritme de MapReduce. Per exemple, resoldre l'interval temporal ZOHE, $S[\tau, \tau + \delta]^{\text{ZOHE}} = S(\tau, \tau + \delta] \cup \{(\tau + \delta, V(\inf(S[\tau + \delta, +\infty)))\}$ (v. def. 4.49), implica conèixer la mesura següent a un $\tau + \delta$ i per tant treballar sobre tota la sèrie temporal original, cosa que no és possible perquè en les etapes map només es treballa sobre un subconjunt de la sèrie temporal original. Això no obstant, per al RoundRobindoop definit, si considerem que la sèrie temporal no té inframostreig aleshores en

l'etapa map podem fer la selecció prèvia per l'interval $S(\tau, \tau + 2\delta]$, és a dir assumim que hi ha mesura a l'interval $[\tau + \delta, \tau + 2\delta]$, i a l'etapa reduce ja es calcularà correctament $f(S, \tau, \delta)$ per a l'interval $[\tau, \tau + \delta]$.

Noteu que en l'algoritme de RoundRobindoop definit hem assumit que l'interval de selecció en l'etapa map sempre és $(\tau, \tau + \delta]$ i en l'etapa reduce usem el model genèric d'agregació $f(S, \tau, \delta)$ tot i que les mesures ja han estat seleccionades. Per tant, la interpretació en l'etapa map no és vàlida per a totes les f.

Per tal d'ampliar l'etapa map, proposem una nova funció de classificació que admeti ampliar l'interval de selecció. Si en la classificació definida cada mesura es classificava en un i només un $\tau + \delta$ per a cada resolució, en la nova funció de classificació es pot escollir a quants $\tau^* + \delta$ es classifica cada mesura. Sigui classifica(m, E) la funció de classificació original i siguin $l, g \in \mathbb{N}$ les quantitats desitjades. La nova funció de classificació és

$$\begin{aligned}&\text{classifica}'(m, E, l, g) = \text{classifica}(m, E) \cup \\ &\cup \{(\delta, f, \tau + (n - \gamma)\delta, t, v) | (\delta, f, \tau + n\delta, t, v) \in \text{classifica}(m, E) \wedge \gamma \in \{1, \ldots, g\}\} \\ &\cup \{(\delta, f, \tau + (n + \lambda)\delta, t, v) | (\delta, f, \tau + n\delta, t, v) \in \text{classifica}(m, E) \wedge \lambda \in \{1, \ldots, l\}\}\end{aligned}$$

El paràmetre l permet classificar una mesura en temps posteriors i el paràmetre g en temps anteriors; si ho observem des del punt de vista de la selecció d'interval $S(\tau - l\delta, \tau + (1 + g)\delta]$, el paràmetre l permet estendre l'interval cap a l'esquerra i g cap a la dreta.

Exemple 12.2 (Classificació d'una mesura per ZOHE)**.** Com ja hem comentat, per a les funcions d'agregació d'atributs de la família ZOHE s'ha d'aproximar l'interval temporal ZOHE a una selecció en l'interval $S(\tau, \tau + 2\delta]$. És a dir, que la funció de classifica ha de retornar dues classificacions per a cada mesura, una amb $\tau + \delta$ i l'altra amb τ. Per tant, per aquesta família $g = 1$ i $l = 0$.

Sigui l'esquema de multiresolució $E = \{(\delta_1 = 2, f_1, \tau_1 = 0, \kappa_1 = 4), (\delta_2 = 5, f_2, \tau_2 = 10, \kappa_2 = 3)\}$ i la mesura $m = (25, 1)$, aquesta és classificada per a consolidar-se en les dues resolucions i en els dos instants següents: classifica$'(m, E, l, g) = \{(2, f_1, \tau^* - g\delta_1, 25, 1), (2, f_1, \tau^*, 25, 1), (5, f_2, \tau^{**} - g\delta_2, 25, 1), (5, f_2, \tau^{**}, 25, 1)\}$ on $\tau^* = 26$, $\tau^* - g\delta_0 = 24$, $\tau^{**} = 25$ i $\tau^{**} - g\delta_1 = 20$.

Mentre que a l'exemple 12.1 es classifica la mesura $(25, 1)$ als instants 26 per a δ_1 i 25 per a δ_2, ara s'afegeixen respectivament els instants 24 i 20. Així, mentre que a l'exemple 12.1 s'interpretaven respectivament els intervals de selecció com a $S(24, 26]$ i $S(20, 25]$, ara s'han ampliat a $S(24, 28]$ i $S(22, 26]$ per a δ_1 i $S(20, 30]$ i $S(15, 25]$ per a δ_2.

En resum, el model de programació MapReduce limita les capacitats dels SGSTM, sobretot pel que fa a les funcions d'agregació d'atributs. A continuació, en l'apartat d'execució de l'algoritme, utilitzarem un exemple amb aquests processos de classificació.

Quant a l'ajustament de l'inici de la consolidació

L'algoritme de MapReduce definit no té en compte els cardinals màxims. És a dir, en l'etapa map es classifiquen les mesures en l'interval de consolidació que correspon segons $\tau + n\delta$, per $n \in \mathbb{Z}$. Aleshores, un cop computat el resultat, el cardinal màxim κ decidiria quins intervals de consolidació s'han d'eliminar. Però com que és una computació en diferit, ja es poden conèixer els intervals que hauran de ser eliminats posteriorment i no caldria tenir-los en compte. Aquest problema es pot solucionar fàcilment ajustant els τ inicials de l'esquema multiresolució.

Així, en la computació en diferit, sigui S una sèrie temporal i sigui E un esquema de multiresolució, el temps de la darrera mesura $T(\max S)$ indueix els intervals de consolidació quan es tenen en compte els cardinals màxims. Per a tenir en compte els cardinals màxims s'haurà d'usar un nou esquema E' en què cada τ original sigui canviat a un altre temps de consolidació múltiple $\tau' = \tau + n\delta$, per $n \in \mathbb{Z}$, que tingui en compte $T(\max S)$.

Definició 12.3 (Ajustament del temps d'inici segons el darrer temps de la sèrie temporal)**.** *Sigui $(\delta, f, \tau, \kappa)$ un dels paràmetres de l'esquema de multiresolució i sigui S una sèrie temporal de la qual es coneix $T(\max S)$. Es calcula un nou temps d'inici $\tau' = \tau + n\delta$ (1), on $n \in \mathbb{Z}$, que tingui en compte el cardinal màxim per a $T(\max S)$. És a dir que compleixi $\tau' + \kappa\delta \leq T(\max S)$ (2) i alhora $\tau' + (\kappa+1)\delta \leq T(\max S)$ (3). Això significa que les mesures entre $[\tau', \tau' + \kappa\delta]$ són les pendents de consolidar, les mesures entre $[\tau' + \kappa\delta, T(\max S)]$ encara no es poden consolidar i les anteriors a τ' ja es poden eliminar.* Nota: *sense tenir en compte retards de buffer.*

Càlcul de n: *Substituint la (1) a la (2) s'obté $\tau + n\delta + \kappa\delta \leq T(\max S)$, operant s'obté $\tau + (n+\kappa)\delta \leq T(\max S)$ i $n \leq \frac{T(\max S)-\tau}{\delta} - \kappa$ d'on amb (3) es conclou que $n = \left\lfloor \frac{T(\max S)-\tau}{\delta} - \kappa \right\rfloor$. A més a més, si no es considera vàlid $\tau' < \tau$, és a dir que no es volen mesures abans del temps d'inici original, aleshores n com a mínim pot valdre zero.*

Exemple 12.3 (Classificació d'una mesura en les resolucions amb ajustament dels temps d'inici)**.** Reprenent l'exemple 12.3 per a classificar la mesura $m = (25, 1)$ segons l'esquema de multiresolució $E = \{(\delta_1 = 2, f_1, \tau_1 = 0, \kappa_1 = 4), (\delta_2 = 5, f_2, \tau_2 = 10, \kappa_2 = 3)\}$, ara es volen tenir en compte els cardinals màxims.

Sigui $T(\max S) = 35$ el temps de la mesura màxima de la sèrie temporal original. Aleshores cal canviar l'esquema de multiresolució a $E' = \{(\delta_1 = 2, f_1, \tau_1', \kappa_1 = 4), (\delta_2 = 5, f_2, \tau_2', \kappa_2 = 3)\}$ on $\tau_1' = 26$ i $\tau_2' = 20$ aplicant la definició 12.3. La classificació esdevé classifica$(m, E') = \{(5, f_2, 25, 25, 1)\}$ on no hi ha la resolució δ_1 perquè $25 < \tau_1'$.

Com que l'ajustament dels temps d'inici es pot aplicar per a qualsevol computació en diferit de la multiresolució, implementem un nou mètode a RoundRobinson que

aplica aquest càlcul. L'anomenem set_tau_tnow. És a dir, a partir d'un temps *tnow*, que indica el temps actual en què es troba la sèrie temporal, M.set_tau_tnow(tnow) ajusta tots els temps d'inici de M per tal que quan s'executi la consolidació s'ignorin els intervals que serien immediatament eliminats a causa del cardinal màxim del disc. Al llistat 12.1 reprenem l'exemple del llistat 11.4 per a il·lustrar-ho.

Llistat 12.1: Exemple d'ajustament dels temps d'inici amb RoundRobinson

```
#Importació dels objectes necessaris
>>> from pytsms import TimeSeries, Measure as m
>>> from roundrobinson import MultiresolutionSeries
>>> from roundrobinson.aggregators import mean_zohe,maximum_zohe

#Definició de la sèrie temporal d'exemple
>>> s = TimeSeries([m(1,6),m(5,2),m(8,5),m(10,0),m(14,1),m(19,6),m(22,11),m(26,6),m(29,0)])

#Definició de la sèrie temporal multiresolució
>>> M = MultiresolutionSeries()
#Definició de l'esquema multiresolució
>>> M.addResolution(delta=5,k=4,f=mean_zohe,tau=0)
>>> M.addResolution(delta=10,k=3,f=maximum_zohe,tau=0)

#Ajustament dels temps d'inici segons tnow=29
>>> M.set_tau_tnow(29)
#Consulta dels nous temps d'inici
>>> M.str_taus()
'5/mean_zohe:5 | 10/maximum_zohe:-10'
```

Ara es podria aplicar la consolidació com al llistat 11.4. A diferència, però, aquesta es duria a terme directament en els instants $10, 15, 20, 25$ per a la resolució $\delta = 5$, és a dir sense calcular l'instant 5 i posteriorment eliminar-lo com havíem fet al llistat 11.4. I es duria a terme en $0, 10, 20$ per a $\delta = 10$, en aquest cas no es té en compte cap limitació i per tant es permet un $\tau' = -10$ inferior a $\tau = 0$.

12.2.2. Execució de l'algoritme

Hadoop s'encarrega de l'execució de l'algoritme de MapReduce i de la gestió de les dades d'entrada i de sortida. Per a implementar-lo cal dissenyar un programa per al map i un programa per al reduce, els quals reben de Hadoop els subconjunts de dades escaients i han de retornar els subconjunts també escaients.

Es pot utilitzar diferents llenguatges de programació a l'hora d'implementar l'algoritme de MapReduce, hem escollit el llenguatge Python [106]. Implementem les dues etapes de l'algoritme de MapReduce que hem definit, RoundRobindoop, en un mateix programa que anomenem `rrdoop.py`. El programa té un paràmetre que permet escollir l'etapa: `rrdoop.py -map` o `rrdoop.py -reduce`. A més també hi ha un paràmetre per a definir l'esquema de multiresolució utilitzat, `rrdoop.py -\ map -schema e`.

El programa es comunica amb Hadoop mitjançant l'entrada estàndard (que abreviem amb *stdin*) per a rebre dades i mitjançant la sortida estàndard (que abreviem amb *stdout*) per a retornar els resultats. Gràcies a la generalització del programa amb comunicació per stdin i stdout, també es pot executar l'algoritme de MapReduce al shell del sistema operatiu, cosa que facilita l'experimentació amb l'algoritme. Així, a continuació, primer mostrem l'execució pas a pas de `rrdoop.py` al shell i després mostrem l'execució a Hadoop.

Execució a la shell

El llistat 12.2 és l'execució de `rrdoop.py` al shell del sistema operatiu. Només hi ha un procés map i un procés reduce. Es comuniquen les dades a través de pipes (`|`) de la shell i d'un procés d'ordenació (`sort`) que emula el procés d'ordenació per identificador que faria Hadoop. A més, també s'emula el procés de lectura (`cat`) de les dades originals.

Llistat 12.2: Execució a la shell de rrdoop.py

```
cat original.csv | rrdoop.py -map -schema e.pickle -mapg 1 | sort\
 -k1,1 | rrdoop.py -reduce -schema e.pickle  > final.csv
```

Les dades d'entrada són un fitxer, que també podrien ser fitxers de dades, de les quals Hadoop en processa conjunts de línies a cada procés map. Aquestes dades no cal que siguin ordenades i cal tenir en compte que es poden trencar per qualsevol línia, tot i que Hadoop permet configurar etapes que defineixin com s'han de partir els fitxers. Al llistat 12.3 mostrem les dades d'entrada emmagatzemades en el fitxer `original.csv`, que es corresponen amb la sèrie temporal ja utilitzada al llistat 11.4. Aquest fitxer de dades té format de CSV, com ja s'ha vist al llistat 11.2 en les funcionalitats complementàries per a l'emmagatzematge de Pytsms.

Llistat 12.3: Dades d'entrada original.csv

```
1,6
8,5
5,2
10,0
14,1
19,6
26,6
29,0
22,11
```

El fitxer `original.csv` es transmet a través de `cat` i pipe a l'stdin del procés de map `rrdoop.py -map -schema e.pickle -mapg 1`. El procés de map té un esquema de multiresolució com a paràmetre, `rrdoop.py` a través del paràmetre

`-schema` admet una sèrie temporal multiresolució en format Pickle, com s'ha vist al llistat 11.5, l'esquema de la qual serà el que s'utilitzi. En aquest cas, `e.pickle` es correspon amb el fitxer `mrd.pickle` del llistat 11.5 i per tant amb l'esquema de multiresolució del llistat 11.4: $E = \{(\delta_0 = 5, \kappa_0 = 4, f_0 = \text{mitjana}^{\text{ZOHE}}, \tau_0 = 0), (\delta_1 = 10, \kappa_1 = 3, f_1 = \text{màxim}^{\text{ZOHE}}, \tau_1 = 0)\}$.

En aquest exemple d'esquema de multiresolució s'usen funcions d'agregació d'atributs de la família ZOHE. Com ja hem comentat a l'exemple 12.2, s'ha de canviar la selecció que l'etapa map duu a terme. A tal efecte RoundRobindoop admet un paràmetre `-mapg` per a indicar l'expansió de la classificació cap a la dreta. Aproximem l'interval temporal ZOHE a una selecció en l'interval $(\tau, \tau + 2\delta]$, és a dir que hem d'expandir un interval `-mapg 1`. RoundRobindoop també admet un paràmetre `-mapl` per a l'expansió cap a l'esquerra, que en aquest cas no cal utilitzar.

El procés de map retorna el resulta a través de l'stdout, el qual es mostra al llistat 12.4 i es correspon amb les dades d'entremig de MapReduce. El format és el requerit per Hadoop, és a dir cada línia és una parella d'identificador i valor separats per un tabulador. L'identificador és $(\delta, f, \tau + \delta)$ però escrit en el format $\delta/f\text{–}\tau + \delta$ i el valor és (t, v) escrit separat per un espai. Es pot observar com es comença classificant la primera mesura $(1, 6)$ en l'instant de consolidació 10 per δ_1 i 5 per δ_0. A continuació la mesura $(8, 5)$ es classifica a l'instant de consolidació 10 per δ_1 i en els 10 i 5 per δ_0, en aquest darrer cas s'aplica l'aproximació de la selecció ZOHE per l'interval $(\tau, \tau + 2\delta]$; cal destacar que en els casos anteriors no s'aplica perquè resultaria en un $\tau' <= \tau$. I així per a totes fins a la darrera mesura $(22, 11)$.

Llistat 12.4: Sortida del procés map

```
10/maximum_zohe-10      1 6.0
5/mean_zohe-5   1 6.0
10/maximum_zohe-10      8 5.0
5/mean_zohe-10  8 5.0
5/mean_zohe-5   8 5.0
10/maximum_zohe-10      5 2.0
5/mean_zohe-5   5 2.0
10/maximum_zohe-10      10 0.0
5/mean_zohe-10  10 0.0
5/mean_zohe-5   10 0.0
10/maximum_zohe-20      14 1.0
10/maximum_zohe-10      14 1.0
5/mean_zohe-15  14 1.0
5/mean_zohe-10  14 1.0
10/maximum_zohe-20      19 6.0
10/maximum_zohe-10      19 6.0
5/mean_zohe-20  19 6.0
5/mean_zohe-15  19 6.0
```

```
10/maximum_zohe-30      26 6.0
10/maximum_zohe-20      26 6.0
5/mean_zohe-30  26 6.0
5/mean_zohe-25  26 6.0
10/maximum_zohe-30      29 0.0
10/maximum_zohe-20      29 0.0
5/mean_zohe-30  29 0.0
5/mean_zohe-25  29 0.0
10/maximum_zohe-30      22 11.0
10/maximum_zohe-20      22 11.0
5/mean_zohe-25  22 11.0
5/mean_zohe-20  22 11.0
```

A continuació el procés d'ordenació `sort -k1,1` ordena per identificadors, el qual es mostra al llistat 12.5. Hadoop agruparia els mateixos identificadors i els transmetria a l'stdin d'un procés de reduce. Observem per exemple la primera resolució `10/\ maximum_zohe-10` que conté les mesures en els instants de temps 10, 14, 1 ,19, 5 i 8; l'agregació posterior haurà de treballar en l'interval ZOHE [0,10] i per tant ara queda clar que les mesures de 14 i 19 no són necessàries, però això no ho podíem resoldre en l'etapa de map.

Llistat 12.5: Sortida del procés d'ordenació

```
10/maximum_zohe-10      10 0.0
10/maximum_zohe-10      14 1.0
10/maximum_zohe-10      1 6.0
10/maximum_zohe-10      19 6.0
10/maximum_zohe-10      5 2.0
10/maximum_zohe-10      8 5.0
10/maximum_zohe-20      14 1.0
10/maximum_zohe-20      19 6.0
10/maximum_zohe-20      22 11.0
10/maximum_zohe-20      26 6.0
10/maximum_zohe-20      29 0.0
10/maximum_zohe-30      22 11.0
10/maximum_zohe-30      26 6.0
10/maximum_zohe-30      29 0.0
5/mean_zohe-10  10 0.0
5/mean_zohe-10  14 1.0
5/mean_zohe-10  8 5.0
5/mean_zohe-15  14 1.0
5/mean_zohe-15  19 6.0
5/mean_zohe-20  19 6.0
5/mean_zohe-20  22 11.0
5/mean_zohe-25  22 11.0
```

```
5/mean_zohe-25   26  6.0
5/mean_zohe-25   29  0.0
5/mean_zohe-30   26  6.0
5/mean_zohe-30   29  0.0
5/mean_zohe-5    10  0.0
5/mean_zohe-5    1  6.0
5/mean_zohe-5    5  2.0
5/mean_zohe-5    8  5.0
```

Finalment, el procés de reduce `rrdoop.py -reduce -schema e.pickle > final\ .csv` obté de l'stdin les dades del llistat 12.5 i retorna les dades finals per l'stdout que està redirigit al fitxer `final.csv`, el contingut del qual es mostra al llistat 12.6. Hadoop emmagatzemaria aquestes dades en un fitxer o fitxers de dades. Aquestes dades tenen el format $\delta\ f\ \tau + \delta\ v$ on $(\tau + \delta, v)$ és la mesura consolidada per a la resolució identificada per (δ, f).

Llistat 12.6: Dades de sortida final.csv

```
10 maximum_zohe 10 6.0
10 maximum_zohe 20 11.0
10 maximum_zohe 30 None
5 mean_zohe     10 3.0
5 mean_zohe     15 2.0
5 mean_zohe     20 7.0
5 mean_zohe     25 8.0
5 mean_zohe     30 None
5 mean_zohe     5 2.8
```

Així aquest resultat és el mateix que el de les consultes SèrieDisc$(M, 5, \text{mitjana}^{\text{ZOHE}})$ i SèrieDisc$(M, 10, \text{màxim}^{\text{ZOHE}})$ del llistat 11.4 però amb les particularitats següents:

- RoundRobindoop no té en compte els cardinals màxims κ de les resolucions. Hi ha la mesura consolidada a l'instant 5 per a la resolució $\delta_0 = 5$ que ja hauria d'haver estat eliminada per complir amb $\kappa_0 = 4$.
- RoundRobindoop no té en compte el temps màxim de la sèrie temporal original per a conèixer les mesures que encara no són consolidables. Hi ha la mesures en l'instant 30 per a $\delta_0 = 5$ i $\delta_1 = 10$ que encara no podien ser calculades perquè $T(\max S) = 29$. De fet tenen valor nul (*None*) a causa que l'interval ZOHE no es pot calcular.
- Així doncs, hi ha 9 mesures consolidades finals però 3 s'han de descartar. Un cop s'hagin descartat, per a la resolució $\delta_0 = 5$ hi haurà 4 mesures i es complirà $4 \leq \kappa_0 = 4$, i per a la resolució $\delta_1 = 10$ hi haurà 2 mesures i es complirà $2 \leq \kappa_1 = 3$.

El fitxer del llistat 12.6 té el format CSV de RoundRobinson. Per tant, es poden recuperar aquestes dades amb les operacions de `LoadCsv`, com s'ha descrit al llistat 11.5, i aplicar-hi les operacions de consulta convenients: obtenir la sèrie temporal total, obtenir les subsèries, fer-ne gràfics, etc.

Execució a Hadoop

El llistat 12.7 mostra els passos d'execució de `rrdoop.py` a Hadoop. Primer cal copiar la sèrie temporal original a HDFS, després s'executa l'algoritme MapReduce i finalment es recupera el resultat de HDFS. Hadoop *streaming* és l'eina que permet l'execució a Hadoop de qualsevol programa, en qualsevol llenguatge, que tingui el model de MapReduce. Per a més detall sobre les ordres i els processos vegeu la documentació de Hadoop [124], en aquest exemple hem simplificat amb [...] algunes adreces.

Llistat 12.7: Execució a Hadoop de rrdoop.py

```
hadoop dfs -copyFromLocal original.csv [...]original.csv

hadoop jar [...]hadoop-streaming*.jar -file rrdoop.py -file e.\
pickle  -mapper 'rrdoop.py -map -schema e.pickle -mapg 1' -\
reducer 'rrdoop.py -reduce -schema e.pickle' -input [...]original\
.csv -output [...]final

hadoop dfs -copyToLocal [...]final/part-00000 final.csv
```

Per tal d'observar de manera senzilla l'execució de l'algoritme a Hadoop hem realitzat una configuració anomenada *Single Node Setup*, és a dir on només hi ha un computador que processa. Un cop verificat es podria estendre a una configuració de *Cluster Setup*, en què hi hagués més computadors on distribuir les dades i els processos.

El resultat és el mateix fitxer `final.csv` que per a l'execució a la shell, és a dir el llistat 12.6. Hadoop gestiona automàticament la distribució i la quantitat dels processos map i reduce. Així, alguns dels map o reduces es poden ajuntar en el mateix procés, per exemple els reduce poden rebre tant un subconjunt de les dades d'entremig amb el mateix identificador com un conjunt d'aquests subconjunts, però mai se separa el mateix identificador en diferents reduces. De fet, en el cas anterior que hem executat a la shell s'utilitza un conjunt amb tots els subconjunts possibles, ja que només hi ha un procés map i un de reduce.

13. Implementació relacional amb Tutorial D

Els models SGST i SGSTM es basen fortament en l'àlgebra relacional. Així doncs, n'explorem una implementació en un SGBDR. Per tal de mantenir la màxima fidelitat amb els models i per les consideracions de Date [23, cap. 1–4] sobre la idoneïtat del llenguatge SQL en el model relacional, implementem els models amb el llenguatge acadèmic dels SGBDR: *Tutorial D* [23, 30, 32]. Aquest llenguatge és orientat a l'àlgebra relacional i per tant és còmode per a implementar els models que hem definit fortament basats en aquesta àlgebra. Com a intèrpret per a aquest llenguatge utilitzem Rel [129], el qual és un SGBDR de propòsit acadèmic per a experimentar amb Tutorial D.

Implementem la part essencial, és a dir l'àlgebra definida en els models lògics. Com que són de propòsit acadèmic, ni Tutorial D ni Rel implementen els complements que tenen els SGBDR en entorns d'explotació. Com a conseqüència, les implementacions resultants són útils per a comprovar l'àlgebra relacional però no són còmodes per a treballar amb dades reals, les quals necessiten sovint operacions per canviar-ne el format, per dibuixar-les, etc.

En primer lloc, implementem un SGST relacional anomenat *Reltsms*. En segon lloc, pel cas dels SGSTM, implementem la funció de multiresolució sobre Reltsms.

13.1. Reltsms

Reltsms és una col·lecció d'operadors definits amb Tutorial D sobre una estructura de sèrie temporal definida com a relació. A continuació en destaquem algunes definicions i alguns casos d'ús. La sintaxi de Tutorial D utilitzada es correspon amb l'acceptada per Rel [129].

Tal com s'ha definit el model estructural de SGST, en el model relacional les sèries temporals són relacions amb dos atributs: t pel temps i v pels valors, on l'atribut t fa de clau primària en les variables relació. Les mesures són els tuples d'aquestes relacions. Per tant, la implementació relacional consisteix a definir un tipus sèrie temporal i totes les operacions associades. Aquestes definicions es basaran en Tutorial D, és a dir que no cal cap llenguatge no relacional extern.

Tutorial D defineix estàticament els tipus dels paràmetres de les operacions, la qual cosa dificulta generalitzar-les per a qualsevol tipus de temps i valor com s'ha fet a la implementació amb Python. Com a conseqüència, fixarem a reals els tipus dels atributs temps i dels valors, *Rational* a Tutorial D.

Així doncs, definim el valor de sèrie temporal mitjançant una relació d'atributs t i v. Per exemple la sèrie temporal $s = \{(2, 3), (3, 4)\}$:

```
relation {
   tuple { t 2.0, v 3.0 },
   tuple { t 3.0, v 4.0 }
}
```

```
  t      v
———— ————
2.0  3.0
3.0  4.0
```

A partir d'aquest valors relació s'hauria de definir el tipus sèrie temporal. No obstant això, la definició de tipus es troba en un estat massa experimental a Tutorial D, com es detalla a l'apartat 13.1.2. Així doncs, definim una variable relació *timeseries* per a utilitzar-la com a equivalent del tipus en les definicions dels operadors:

```
var timeseries base relation
    { t rational, v rational } key { t } ;
```

En les sèries temporals com a relació, els tuples són les mesures del model de SGST. Hi ha alguns operadors dels SGST que tenen mesures com a paràmetres, així doncs caldria definir també un tipus per a les mesures. Per a simplificar, definim aquestes mesures com a sèries temporals d'un sol element, per exemple la mesura $m = (2, 3)$:

```
relation {
   tuple { t 2.0, v 3.0 }
}
```

A partir d'aquestes mesures com a valors relació definim els operadors que tenen mesures com a paràmetres. Així, els dos operadors bàsics dels SGST per a obtenir els atributs de temps i de valor d'una mesura són:

```
operator ts.t(m same_type_as (timeseries)) returns rational;
  return t from tuple from m;
end operator;

operator ts.v(m same_type_as (timeseries)) returns rational;
  return v from tuple from m;
end operator;
```

Per exemple, per a obtenir l'atribut temps $T(m)$ de la mesura:

```
with relation {
    tuple { t 2.0, v 3.0 }
  } as m1:
ts.t(m1)
```

```
2.0
```

i per a obtenir el valor $V(m)$:

```
with relation {
    tuple { t 2.0, v 3.0 }
  } as m1:
ts.v(m1)
```

```
3.0
```

13.1.1. Operadors

Un cop implementat el model estructural és el torn d'implementar el model d'operacions. Algunes operacions dels SGST es poden implementar directament amb les operacions relacionals de Tutorial D aplicades a les relacions *timeseries*, és el cas de:

- La projecció

```
with
 relation {
    tuple { t 2.0, v 3.0 },
    tuple { t 3.0, v 4.0 }
 } as s:
s {t}
```

```
  t
-----
2.0
3.0
```

- La selecció

```
with
 relation {
    tuple { t 2.0, v 3.0 },
    tuple { t 3.0, v 4.0 }
 } as s:
s where v>3.0
```

```
  t     v
-----  -----
3.0    4.0
```

- El reanomena

```
with
 relation {
   tuple { t 2.0, v 3.0 },
   tuple { t 3.0, v 4.0 }
 } as s:
s rename ( v as temperatura )
```

```
 t   temperatura
———  ————————————
3.0            4.0
```

- El mapa

```
with
 relation {
   tuple { t 2.0, v 3.0 },
   tuple { t 3.0, v 4.0 }
 } as s:
extend s add ( t+v as vp)
```

```
 t    v   vp
———  ———  ———
2.0  5.0  5.0
3.0  7.0  7.0
```

- L'agregació, per algunes funcions

```
with
 relation {
   tuple { t 2.0, v 3.0 },
   tuple { t 3.0, v 4.0 }
 } as s:
summarize s add (max(t) as t, sum(v) as v )
```

```
 t    v
———  ———
3.0  1.0
```

La resta d'operacions s'han de definir, les implementem basades en altres operadors de l'àlgebra relacional de Tutorial D.. A continuació en mostrem algunes d'exemple, tots els operadors que definim els anomenem prefixats amb 'ts.'.

L'operació d'unió és equivalent a unir la primera sèrie temporal amb les mesures de la segona que no tenen un atribut de temps igual a algun temps de la primera. Així, l'operador d'unió *ts.union* és:

```
operator ts.union(
        s1 same_type_as (timeseries),
        s2 same_type_as (timeseries))
        returns relation same_heading_as (timeseries);
  return s1 union (s2 join (s2 {t} minus s1 {t}));
end operator;
```

Per exemple, siguin dues sèries temporals $s_1 = \{(2,3),(4,2),(6,4)\}$ i $s_2 = \{(1,2),(5,3),(6,5),(10,1)\}$ la unió $s_1 \cup s_2$ és:

```
with
 relation {
    tuple { t 2.0, v 3.0 },
    tuple { t 4.0, v 2.0 },
    tuple { t 6.0, v 4.0 }
  } as s1,
 relation {
    tuple { t 1.0, v 2.0 },
    tuple { t 5.0, v 3.0 },
    tuple { t 6.0, v 5.0 },
    tuple { t 9.0, v 1.0 }
  } as s2:
ts.union(s1,s2)
```

t	v
1.0	2.0
2.0	3.0
4.0	2.0
5.0	3.0
6.0	4.0
9.0	1.0

Les operacions de funció temporal operen en base a un mètode de representació de les sèries temporals. Els operadors de funció temporal definits en el model de SGST depenen de l'operador d'interval temporal, per tant cal definir-ne un per a cada representació i els altres operadors utilitzaran el que correspongui a la representació sol · licitada.

Així doncs, com a exemple definim l'operador d'interval temporal ZOHE *ts.interval.zohe*:

```
operator ts.interval.zohe(
          s same_type_as (timeseries),
          l rational,
          h rational )
          returns relation same_heading_as (timeseries);
  begin;
    var x rational init(0.0);
    var sp private same_type_as (timeseries) key { t };
    x := ts.v(ts.inf(ts.interval.closed(s,h,1.0/0.0)));
    sp := relation {
```

```
      tuple {t h, v x}
      };
    return ts.union(ts.interval.left(s,l,h),sp);
  end;
end operator;
```

Per exemple, l'interval temporal ZOHE $s_1[1,3]^{\text{ZOHE}}$ és:

```
with
 relation {
   tuple { t 2.0, v 3.0 },
   tuple { t 4.0, v 2.0 },
   tuple { t 6.0, v 4.0 }
  } as s1:
ts.interval.zohe(s1,1.0,3.0)
```

t	v
2.0	3.0
3.0	2.0

Els altres operadors de funció temporal tenen un paràmetre *repr* a partir del qual s'executa l'operador d'interval temporal corresponent, el qual ha d'estar definit prèviament a l'execució. Ni Tutorial D ni Rel ofereixen una manera còmode d'executar un operador escollit segons un paràmetre tot i que Voorhis [128] està desenvolupant millores en aquest sentit. Una solució actualment és l'execució d'una cadena de text mitjançant la sentència *execute*. A continuació ho mostrem en la implementació de l'operació de concatenació temporal *ts.concatenation.t*:

```
operator ts.concatenation.t(
          s1 same_type_as (timeseries),
          s2 same_type_as (timeseries),
          repr char)
          returns relation same_heading_as (timeseries);
  begin;
    var t1 rational init(0.0);
    var t2 rational init(0.0);
    var exp char init('');

    t1 := ts.t(ts.inf(s1));
    t2 := ts.t(ts.sup(s1));

    //Cal executar: ts.globalexecute := ts.interval.<repr>(s2,t1,t2);
    ts.globalexecute := s2;
    exp := 'ts.globalexecute := ts.interval.' || repr || '(ts.globalexecute,' || t1 || ',' || t2 || ');';
    execute exp;

    return ts.union(s1,s2 minus ts.globalexecute);
  end;
end operator;
```

Per exemple, la concatenació ZOHE de les sèries temporals $s1||^{\text{ZOHE}}s2$ és:

```
with
 relation {
   tuple { t 2.0, v 3.0 },
   tuple { t 4.0, v 2.0 },
   tuple { t 6.0, v 4.0 }
  } as s1,
 relation {
   tuple { t 1.0, v 2.0 },
   tuple { t 5.0, v 3.0 },
   tuple { t 6.0, v 5.0 },
   tuple { t 9.0, v 1.0 }
  } as s2:
ts.concatenation.t(s1,s2,'zohe')
```

t	v
1.0	2.0
2.0	3.0
4.0	2.0
6.0	4.0
9.0	1.0

13.1.2. Quant a definir el tipus sèrie temporal

Tutorial D permet definir nous tipus de dades, que poden ser usats amb la mateixa funcionalitat que els tipus predefinits –*char*, *rational*, etc. A l'hora de definir nous tipus, Tutorial D usa el concepte de possibles representacions (*possrep*) del model relacional. Les possibles representacions són les diverses maneres d'instanciar el valor d'un tipus.

Així doncs, en comptes de definir una variable *timeseries* com hem fet en el Reltsms seria millor definir un tipus *timeseries* que tingui una possible representació *ts* que és la relació de sèrie temporal:

```
type timeseries
  possrep {ts relation {t rational, v rational}};
```

Aleshores es poden instanciar sèries temporals:

```
timeseries(
 relation {
   tuple { t 2.0, v 3.0 },
   tuple { t 3.0, v 4.0 }
 }
)
```

de les quals es pot demanar la relació continguda en la representació *ts*:

```
with
 timeseries(
  relation {
    tuple { t 2.0, v 3.0 },
    tuple { t 3.0, v 4.0 }
  }
 ) as s:
THE_ts(s)
```

```
  t     v
 ———   ———
 2.0   3.0
 3.0   4.0
```

Però actualment, la definició de les sèries temporals com a tipus a Tutorial D és incòmode principalment per dos motius:

- Emmascara el concepte de tipus relació per als propòsits acadèmics que tenim. Principalment, Tutorial D no permet definir objectes que siguin del mateix ordre que les relacions.
- La definició estàtica de tipus en les operacions i la rigidesa en les capçaleres de les relacions. Aquest, de fet, és també un inconvenient que ens trobem en la definició dels operadors de Reltsms. Impedeix definir sèries temporals multivaluades i sèries amb atributs de qualssevol tipus.

La sèrie temporal a partir del tipus relació

Els valors relació en l'àlgebra relacional sempre són d'un tipus i aquest tipus es correspon amb la seva capçalera. Però les relacions de Tutorial D són uns objectes de més ordre que la resta de tipus, tant els predefinits com els d'usuari, ja que tenen un tractament molt diferenciat i no es poden declarar nous tipus amb la mateixa funcionalitat que les relacions. Per exemple no es pot heretar del tipus relació i per tant no es poden definir tipus que siguin admesos com a arguments en els operadors predefinits de l'àlgebra relacional.

A Reltsms ens ha sigut còmode definir les sèries temporals com a tipus relació perquè conserva molta correspondència amb el model algebraic. Així en els operadors hem definit els paràmetres de sèrie temporal com de tipus relació d'atributs t i v reals, és a dir en Tutorial D `relation {t rational, v rational}`. De fet hem utilitzat la variable *timeseries* per a agrupar aquesta capçalera i així en els operadors hem definit el tipus a partir de `s same_type_as(timeseries)`.

Si haguéssim definit el tipus *timeseries* aleshores hauríem definit els operadors de la forma següent, per exemple pel cas de la unió a partir de l'operador definit de *ts.union*:

```
operator tstype.union(
        s1 timeseries,
        s2 timeseries,
        returns timeseries;
  return timeseries(ts.union(THE_ts(s1),THE_ts(s2));
end operator;
```

És a dir que *s1* i *s2* són de tipus *timeseries* i per obtenir-ne la relació continguda usem l'operador *THE_ts*. El resultat també és de tipus *timeseries*, el qual s'instancia a partir d'un valor relació.

Aleshores, el mateix exemple que per la unió ara s'executaria amb:

```
tstype.union(timeseries(s1),timeseries(s2))
```

A més, en els usos d'operadors de l'àlgebra relacional també s'hauria d'executar usant l'operador *THE_ts* i tornant a instanciar com a sèrie temporal. Per exemple:

```
with
timeseries(
 relation {
   tuple { t 2.0, v 3.0 },
   tuple { t 3.0, v 4.0 }
  }) as s:
timeseries(THE_ts(s) where v>3.0)
```

En conclusió, definir la sèrie temporal com a tipus *timeseries* en comptes de directament com a tipus relació dificulta poder observar la relació directa entre l'àlgebra relacional del model i la implementació en un SGBDR. Si bé és cert que ambdues es poden usar amb la mateixa funcionalitat, en el cas dels tipus les crides als operadors queden emmascarades amb l'ús de l'operador d'atribut *THE_ts* i la instància de nou de la *timeseries*. A Tutorial D faria falta un sistema d'herència més clar que pogués declarar que el tipus *timeseries* té l'atribut *ts* que és de tipus relació i pot actuar com a tal allà on calgui. Així, per bé que l'emmascarament del tipus podria ser útil en un sistema productiu, dificulta el propòsit acadèmic d'aquesta implementació.

Sèries temporals multivaluades i de qualssevol tipus

Els operadors de Reltsms definits treballen amb la sèrie temporal en forma canònica ja que així és com hem definit la variable *timeseries*, com a relació de dos atributs t i v de tipus real per a simplificar.

Per una banda, es poden definir sèries temporals multivaluades, que són relacions amb l'atribut t i altres atributs per als valors multivaluats. Per exemple la sèrie temporal multivaluada $m = \{(2, 1, 2), (3, 1, 2)\}$:

```
relation {
  tuple { t 2.0, v1 1.0, v2 2.0},
  tuple { t 3.0, v1 1.0, v2 2.0}
 }
```

```
 t      v1    v2
———  ———  ———
2.0  1.0  2.0
3.0  1.0  2.0
```

També es poden definir sèries temporals multivaluades en la forma canònica, que son relacions amb l'atribut t i v on el v és una relació dels atributs multivaluats. Per exemple la sèrie temporal multivaluada m en la forma canònica és $m_c = \{(2, (1, 2)), (3, (1, 2))\}$:

```
relation {
  tuple { t 2.0, v relation { tuple {v1 1.0, v2 2.0 }} },
  tuple { t 3.0, v relation { tuple {v1 1.0, v2 2.0 }} }
 }
```

```
           v
 t     ———————
       v1    v2
———  ———  ———
2.0  1.0  2.0
3.0  1.0  2.0
```

Els operadors d'agrupament de l'àlgebra relacional permeten transformar de la forma multivaluada a la forma canònica i a la inversa. Així, amb Tutorial D l'operació de transformació de la forma multivaluada a la canònica és:

```
s group ({all but t} as v)
```

i la transformació de la forma canònica a la multivaluada és:

```
s ungroup (v)
```

Per altra banda, es poden definir sèries temporals on els atributs no siguin de tipus real sinó per exemple enters:

```
relation {
  tuple { t 2, v 1},
  tuple { t 3, v 1}
 }
```

```
 t      v
———  ———
 2      1
 3      1
```

Això no obstant, a Tutorial D aquestes relacions de sèries temporals multivaluades o d'altres tipus que no siguin reals no són acceptades com a arguments en els operadors definits de Reltsms. Les sèries temporals multivaluades són relacions amb la capçalera $\{t, v1, v2, \ldots\}$ o bé $\{t, v : \{v1, v2, \}\}$ en la forma multivaluada canònica, les quals a Tutorial D no s'adiuen amb la capçalera $\{t : \mathbb{R}, v : \mathbb{R}\}$ de les *timeseries*. El mateix ocorre amb relacions amb capçalera on els tipus no són reals, per exemple on són enters $\{t : \mathbb{Z}, v : \mathbb{Z}\}$.

Això és degut al fet que a Tutorial D hi ha definició estàtica dels tipus i en els operadors de Reltsms els paràmetres són del mateix tipus que *timeseries*, és a dir exactament de tipus `relation {t rational, v rational}`. El mateix problema també ocorre en definir el tipus *timeseries* amb una possible representació d'aquest tipus, aleshores no es poden instanciar sèries temporals on els atributs no siguin de tipus real.

Per tant, necessitaríem definir les sèries temporals com a relacions de tipus on un dels atributs és t el temps i els altres atributs són qualssevol. Darwen [20] proposa una ampliació de Tutorial D en aquest sentit que permetria definir el tipus relació amb asteriscs per a indicar qualsevol nom d'atribut o qualsevol tipus, per exemple el més genèric seria `relation {*}` que indicaria una relació de qualsevol tipus. Cal tenir en compte, però, que per ara només és una proposta.

Així, més específicament, ens permetria definir les sèries temporals en forma canònica com de tipus `relation {t *, v *}` i la forma multivaluada com de tipus `relation {t * ,*}`. Aleshores podríem definir correctament el tipus sèrie temporal amb les dues possibles representacions de canònica i multivaluada:

```
type timeseries
        possrep ts.canonical {ts relation {t *, v * }}}
        possrep ts.multivalued {multivalued relation {t * ,*}}
        init ts.canonical (multivalued:= ts ungroup (v))
             ts.multivalued (ts:= multivalued group ({all but t} as v));
```

Aleshores podríem instanciar sèries temporals amb altres tipus en els atributs, com per exemple enters:

```
ts.canonical(
 relation {
   tuple { t 2, v 1},
   tuple { t 3, v 1}
  }
)
```

o bé instanciar sèries temporals multivaluades, per exemple:

```
ts.multivalued(
 relation {
   tuple { t 2.0, v1 1.0, v2 2.0},
   tuple { t 3.0, v1 1.0, v2 2.0}
  }
)
```

de les quals se'n podria obtenir tant la possible representació en forma multivaluada

```
THE_multivalued(m)
```

com en la forma canònica

```
THE_ts(m)
```

En conclusió, definir el tipus sèrie temporal seria de molta utilitat si les capçaleres de les relacions poguessin ser genèriques, sobretot el concepte de possibles representacions és molt útil per a implementar la forma canònica i la multivaluada.

13.2. Multiresolució relacional

Implementem la funció de PlecMu de la definició 7.2 sobre els SGST amb l'operador *mtsms.multiresolution*. Té dos paràmetres: la sèrie temporal s que és una *timeseries* de Reltsms i l'esquema multiresolució *schema* que és una relació, és a dir un conjunt, d'atributs *delta*, *tau*, f i k:

```
operator mtsms.multiresolution(
        s same_type_as (timeseries),
        schema relation {delta rational, tau rational, f char, k rational})
        returns relation same_heading_as (timeseries);
    [...]
```

Els atributs de l'esquema multiresolució són de tipus real, seguint amb la simplificació de Reltsms, llevat de f que és un nom d'un operador que actua com a funció d'agregació d'atributs i per tant permet que l'usuari se'ls defineixi. Per a calcular la subsèrie temporal de cada disc a partir de cada tuple de paràmetres multiresolució, l'operador de multiresolució depèn d'un operador *mtsms.dmap*:

```
operator mtsms.dmap(
        s same_type_as (timeseries),
        delta rational,
        tau rational,
        f char,
        k rational)
        returns relation same_heading_as (timeseries);
    [...]
```

L'operador *mtsms.dmap* és el que executarà l'operador d'agregació d'atributs corresponent al nom mitjançant la sentència *execute*, de manera similar a com s'ha descrit en el Reltsms. Els operadors d'agregació d'atributs tenen tres paràmetres: la sèrie temporal s, l'instant de temps menor l i l'instant de temps major h. Noteu que en el model hem definit una funció d'agregació d'atributs sobre una sèrie temporal, un instant de consolidació τ i un pas de consolidació δ, mentre que les

implementem sobre una sèrie temporal i directament sobre l'interval de consolidació $[l = \tau, h = \tau + \delta]$.

Així doncs, prèviament a l'execució de la multiresolució cal tenir definit un operador d'agregació d'atributs. Com a exemple, definim l'operador *mtsms.aaf.maxzohe* per a la funció d'agregació màxim$^{\text{ZOHE}}$:

```
operator mtsms.aaf.maxzohe(
          s same_type_as (timeseries),
          l rational,
          h rational)
          returns relation same_heading_as (timeseries);
  return summarize ts.interval.zohe(s,l,h) add (max(t) as t, max(v) as v);
end operator;
```

Per exemple, sigui la sèrie temporal $s = \{(4,1),(5,6),(8,2)\}$ i la sèrie temporal multiresolució amb esquema $E = \{\{\tau : 0, \delta : 2, \kappa : 4, f : \text{màxim}^{\text{ZOHE}}\}, \{\tau : 0, \delta : 5, \kappa : 2, f : \text{màxim}^{\text{ZOHE}}\}\}$, la PlecMu$(s, E)$ és

```
with
 relation {
   tuple { t 4.0, v 1.0 },
   tuple { t 5.0, v 6.0 },
   tuple { t 8.0, v 2.0 }
  } as s,
 relation {
   tuple { delta 2.0, tau 0.0, f 'mtsms.aaf.maxzohe', k 4.0},
   tuple { delta 5.0, tau 0.0, f 'mtsms.aaf.maxzohe', k 2.0}
  } as schema:
mtsms.multiresolution(s,schema)
```

t	v
0.0	1.0
2.0	1.0
4.0	1.0
6.0	6.0
8.0	2.0

13.3. Resum

En resum, mitjançant el llenguatge acadèmic dels SGBDR, Tutorial D, hem implementat un SGST a partir del qual hem pogut implementar les operacions de multiresolució. En aquest capítol hem presentat les operacions més destacables de les implementacions per tal d'experimentar com els models dissenyats s'adapten al model relacional. Si bé no hem implementat un SGSTM, es podria seguir un raonament similar al que hem fet per als SGST.

14. Exemple d'ús complet

En aquest capítol mostrem un exemple de base de dades multiresolució aplicada a una sèrie temporal real amb dades massives. Utilitzem RoundRobinson i RoundRobindoop per a consolidar la multiresolució d'aquestes dades. En aquest cas no fem servir l'operador *mtsms.multiresolution* de Reltsms perquè Rel no abasta aquestes dades massives.

Com que els resultats són els mateixos per a totes les computacions, primer mostrem i discutim els resultats i finalment comparem la resposta de les dues implementacions. Així doncs, seguim els passos següents:

1. Descrivim la sèrie temporal original.
2. Proposem un esquema de multiresolució.
3. Avaluem els resultats de la consolidació de la sèrie temporal amb l'esquema proposat.
4. Comparem les diferents implementacions per a computar l'exemple.

En les computacions hem utilitzat els reals estesos, $\bar{\mathbb{R}}$, com a domini del temps segons l'estàndard d'Hora Unix (v. § 4.1.1). Això no obstant, per a facilitar-ne la comprensió, en els exemples mostrem el domini del temps amb el format de calendari UTC.

14.1. Dades

Les dades provenen d'un sistema de monitoratge de temperatura en una xarxa de sensors distribuïts [1], en aquest cas ens centrem en les dades d'un sensor.

Les dades es mostren a la figura 14.1. És una sèrie temporal adquirida durant un període d'un any i mig, des del 29 d'abril del 2010 fins al 18 d'octubre del 2011. En aquest gràfic, la sèrie temporal es mostra interpolant linealment les mesures, és a dir amb el mètode de representació FOH. En total hi ha 146.709 mesures emmagatzemades de la temperatura que va adquirir el sensor, en graus Kelvin (K), cada 2 minuts tot i que de forma irregular. Tot i així en el gràfic només mostrem 466 punts, escollits mitjançant una delmació amb una resolució d'un dia, ja que de totes maneres una resolució superior és imperceptible. En el gràfic d'aquesta sèrie

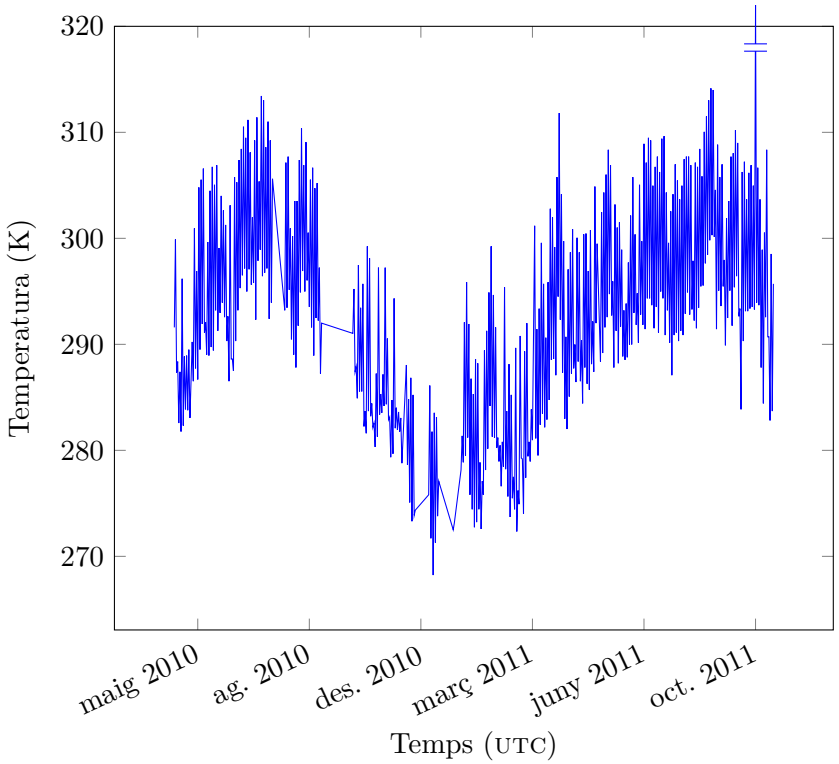


Figura 14.1.: Sèrie temporal d'un sensor de temperatura

temporal destaquen alguns períodes en què hi manquen dades i alguns en què hi ha observacions aberrants.

14.2. Esquema de multiresolució

Dissenyem un esquema de multiresolució per a la sèrie temporal. Aquest esquema resumeix la sèrie temporal amb més resolució per als temps recents i amb menys resolució per als temps antics.

El cronograma de l'esquema de multiresolució es mostra a la figura 14.2, en què cada resolució té un color diferent. Té la mateixa estructura que el cronograma periòdic definit a la figura 5.11 però amb els valors particularitzats per a aquest exemple. De més a menys resolució, l'esquema és el següent:

1. Es consolida una mesura cada 5 hores en un disc amb capacitat de 24 mesures. Per tant, en total aquesta resolució emmagatzema informació durant 5 dies.
2. Es consolida una mesura cada 2 dies en un disc amb capacitat de 20 mesures. Per tant, s'emmagatzema durant 40 dies.

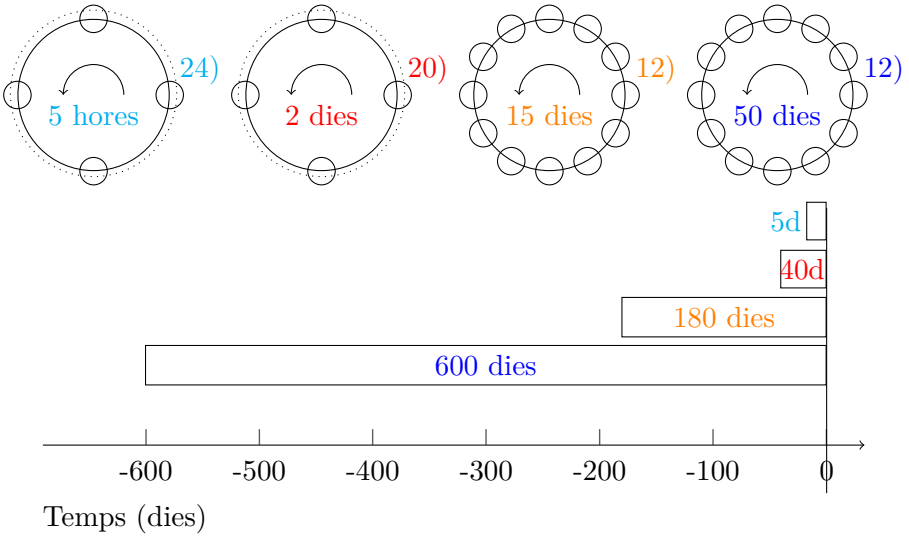


Figura 14.2.: Cronograma de l'esquema de multiresolució

3. Es consolida una mesura cada 15 dies en un disc amb capacitat de 12 mesures. Per tant, s'emmagatzema durant 180 dies.
4. Es consolida una mesura cada 50 dies en un disc amb capacitat de 12 mesures. Per tant, s'emmagatzema durant 600 dies.

Aquest esquema està dissenyat per mostrar diferents paràmetres de la multiresolució. Així el criteri escollit és el de visualitzar algunes resolucions de la sèrie temporal original, la qual té un període de mostreig irregular de 2 minuts, des de la resolució cada 5 hores fins a la resolució cada 50 dies. També s'ha escollit per a mantenir uns lapses de temps repartits, des del lapse de 600 dies que emmagatzema tota la sèrie temporal original fins al de 5 dies que només conté una informació molt recent.

Com a funció d'agregació d'atributs utilitzem la mitjana de la família ZOHE (v. def. 5.38) per a totes les resolucions. A més, per a mostrar diferents maneres d'usar els agregadors, utilitzem el màxim de la família ZOHE però només per a les darreres resolucions.

De manera simplificada, iniciem la consolidació de totes les resolucions al mateix instant de temps, que notem amb τ_0. Com ja hem dit, les dades originals s'inicien el 29 d'abril del 2010, per tant un $\tau_0 = 1$ de gener de 2010 és raonable.

Així doncs, l'expressió de l'esquema de multiresolució en termes de la definició 5.10,

i enumerant cada resolució, és

$$
\begin{aligned}
E = \{ & \\
& (\delta_1 = 5\ \text{h}, f_1 = \text{mitjana}^{\text{ZOHE}}, \kappa_1 = 24, \tau_1 = \tau_0), \\
& (\delta_2 = 2\ \text{d}, f_2 = \text{mitjana}^{\text{ZOHE}}, \kappa_2 = 20, \tau_2 = \tau_0), \\
& (\delta_3 = 15\ \text{d}, f_3 = \text{mitjana}^{\text{ZOHE}}, \kappa_3 = 12, \tau_3 = \tau_0), \\
& (\delta_4 = 50\ \text{d}, f_4 = \text{mitjana}^{\text{ZOHE}}, \kappa_4 = 12, \tau_4 = \tau_0), \\
& (\delta_{3b} = 15\ \text{d}, f_{3b} = \text{màxim}^{\text{ZOHE}}, \kappa_{3b} = 12, \tau_{3b} = \tau_0), \\
& (\delta_{4b} = 50\ \text{d}, f_{4b} = \text{màxim}^{\text{ZOHE}}, \kappa_{4b} = 12, \tau_{4b} = \tau_0), \\
& \}
\end{aligned}
$$

Com que els atributs de mitjana i màxim comparteixen les dues darreres resolucions, les anomenem amb el mateix número però marcant amb una b les de màxim. En total, sumant les capacitats de cada disc, s'emmagatzemen $24 + 20 + 12 + 12 + 12 + 12 = 92$ mesures.

14.3. Resultats de la consolidació

En un SGSTM, consolidem la sèrie temporal original amb l'esquema de multiresolució proposat. Del resultat, podem fer-ne les dues consultes bàsiques (v. § 5.2.3):

1. Les consultes SèrieDisc per a obtenir les subsèries temporals que han quedat consolidades a cada subsèrie resolució.
2. Les consultes SèrieTotal per a obtenir la sèrie temporal total per a cada atribut.

En primer lloc, a la figura 14.3 es mostren totes les subsèries resolució consolidades. Cada gràfic correspon a una consulta possible de SèrieDisc, tot i que les resolucions que només difereixen en l'atribut es mostren en el mateix gràfic. Aquest és el cas de les dues darreres resolucions que comparteixen els mateixos paràmetres llevat de la funció d'agregació d'atributs: en blau es mostra l'atribut de $\text{mitjana}^{\text{ZOHE}}$ i en taronja el de $\text{màxim}^{\text{ZOHE}}$. El títol de cada gràfic indica la subsèrie resolució i els paràmetres de pas de consolidació i de cardinal màxim.

Cada sèrie temporals es mostra amb el mètode de representació ZOHE, car s'han consolidat amb funcions d'agregació basades en aquest mètode. Els eixos de temps marquen els instants de temps però arrodonits en la forma de calendari, per exemple en el primer gràfic hi apareixen les hores però en els altres no encara que cada instant marca una hora concreta. Els valors aberrants es marquen amb discontinuïtats en el gràfic, aquest és el cas de l'atribut màxim del tercer i quart gràfic en què pren un valor de 2938 K. Pel que va a les dades que manquen, les funcions d'agregació

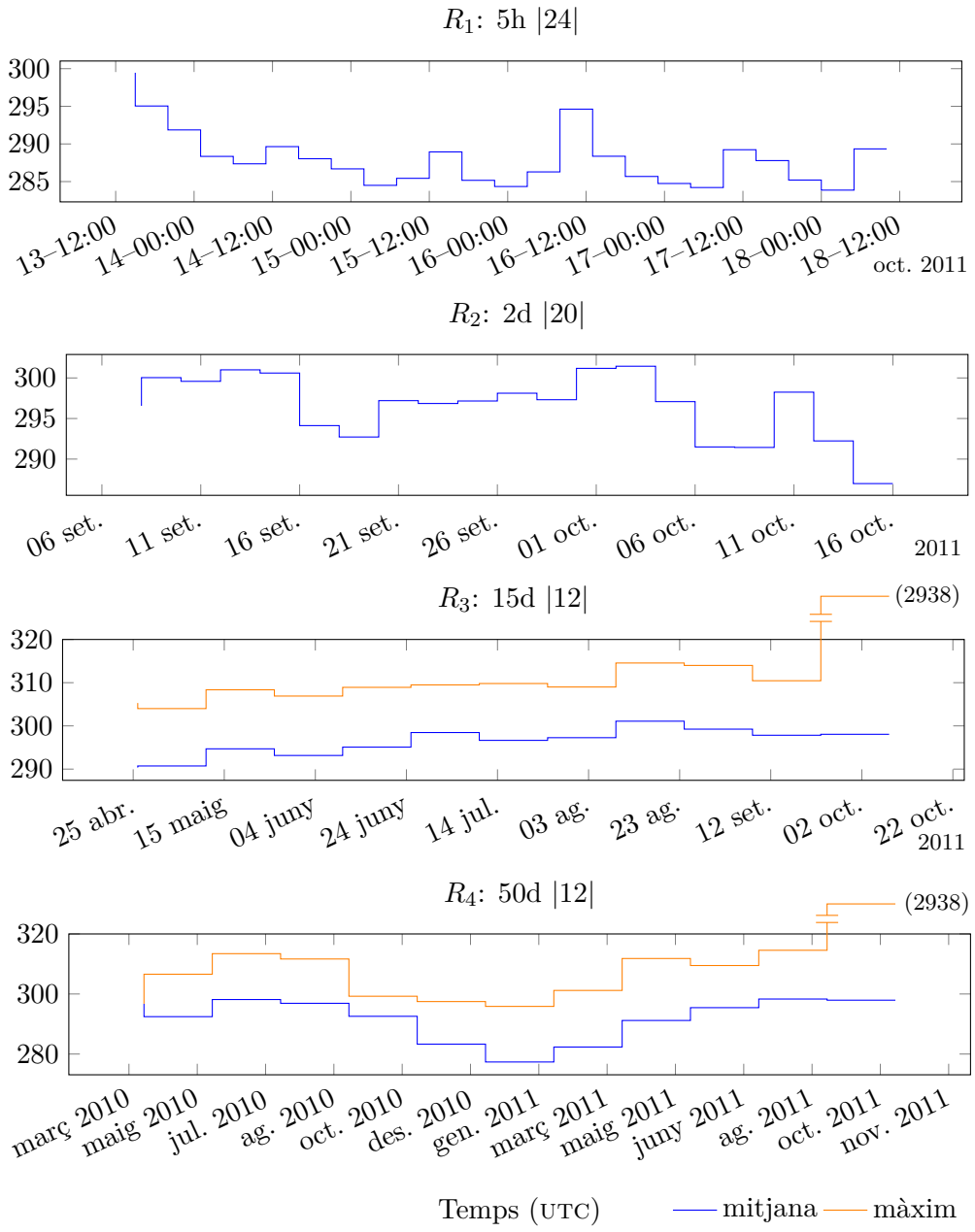


Figura 14.3.: Subsèries resolució emmagatzemades a la base de dades

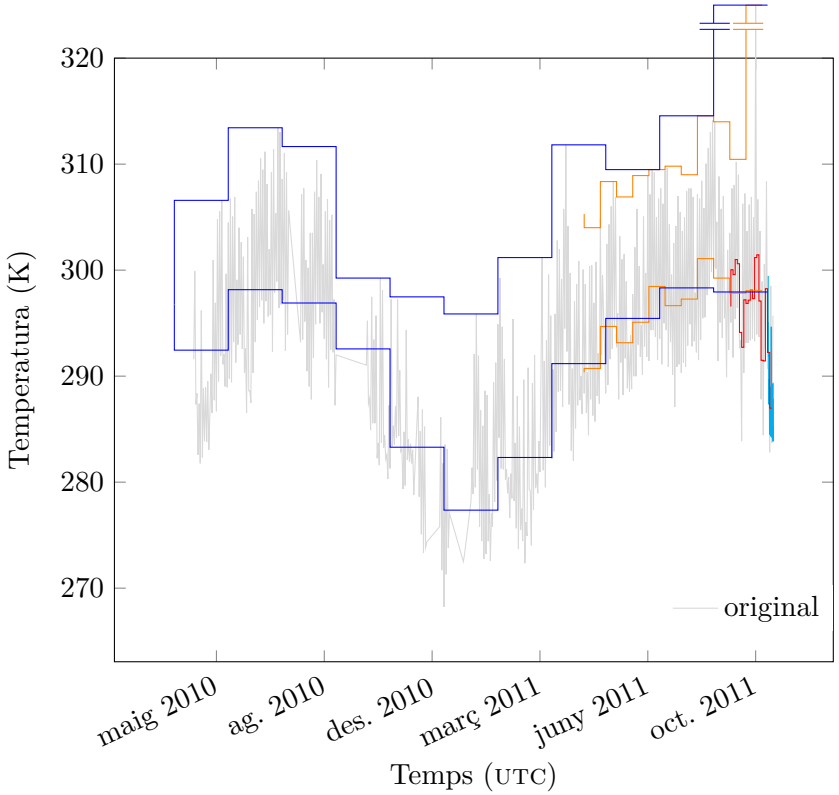


Figura 14.4.: Comparació de la sèrie temporal original, en gris, amb les subsèries resolució, en els mateixos colors que a la figura 14.2

utilitzades han omplert els buits. Això és degut que les agregacions utilitzen el mètode de representació ZOHE i per tant els valors coneguts es mantenen cap enrere, tot i que més correctament per a aquest cas s'hauria de limitar la durada que un valor es pot considerar vàlid per a mantenir-lo o bé s'hauria d'incloure una tècnica de reconstrucció del senyal en les funcions d'agregació d'atributs.

També es pot mostrar totes les subsèries resolució consolidades en un mateix gràfic, com a la figura 14.4 en què cada resolució té el mateix color que a la figura 14.2. En aquest gràfic es pot observar clarament quin tros de la sèrie temporal original (en gris) resumeix cada resolució. Això no obstant, a la figura 14.3 es visualitza més bé la informació de cada resolució, sobretot en el cas de la primera resolució de 5 hores que a la figura 14.4 no se'n distingeixen les mesures.

En segon lloc, a la figura 14.5 es mostra la consulta de la sèrie temporal total per a l'atribut de mitjana$^{\text{ZOHE}}$ (en blau) i per a l'atribut de màxim$^{\text{ZOHE}}$ (en taronja). Ambdues consultes de SèrieTotal s'han calculat amb concatenació temporal ZOHE i es representen gràficament amb el mètode ZOHE. En gris es mostra la sèrie temporal original amb el mètode de representació FOH. Comparant les sèries tem-

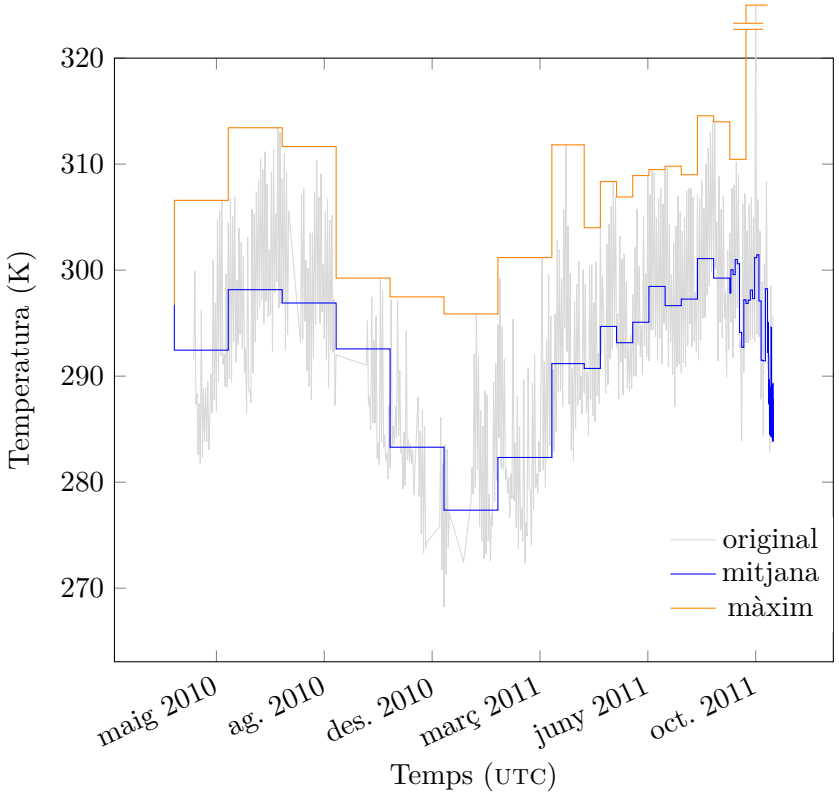


Figura 14.5.: Comparació de la sèrie temporal original amb la sèrie temporal total de la multiresolució per als atributs de mitjana i màxim ZOHE

porals consolidades amb l'original, podem observar que la mitjana s'assembla a un filtre passabaix i el màxim s'assembla a l'envolupant, tot i que calculats segons els períodes de consolidació que marca l'esquema de multiresolució. Així, es pot observar que els atributs tenen una resolució incremental, és a dir menys resolució en els temps antics i més en els temps recents, i que la mitjana acaba tenint més resolució que el màxim.

A tall d'exemple, a la figura 14.6 es representen els atributs també amb el mètode FOH. Com diuen Keogh i Smyth [67], aquesta segmentació de les corbes en línies rectes és més propera a la suavització que realitza la visió humana. Això no obstant, les agregacions estan computades amb ZOHE i per tant la visualització dels atributs sembla desplaçada temporalment cap a la dreta. Per tal que quedés més centrat es podria utilitzar una agregació centrada en l'interval. Tot i així perquè en la representació FOH els màxims quedessin a la cresta de l'envolupant caldria utilitzar una consolidació irregular i més complicada en què el temps de les mesures resultants cerqués els punts que envolupessin el màxim.

En conclusió, aquests resultats exemplifiquen com un SGSTM emmagatzema una

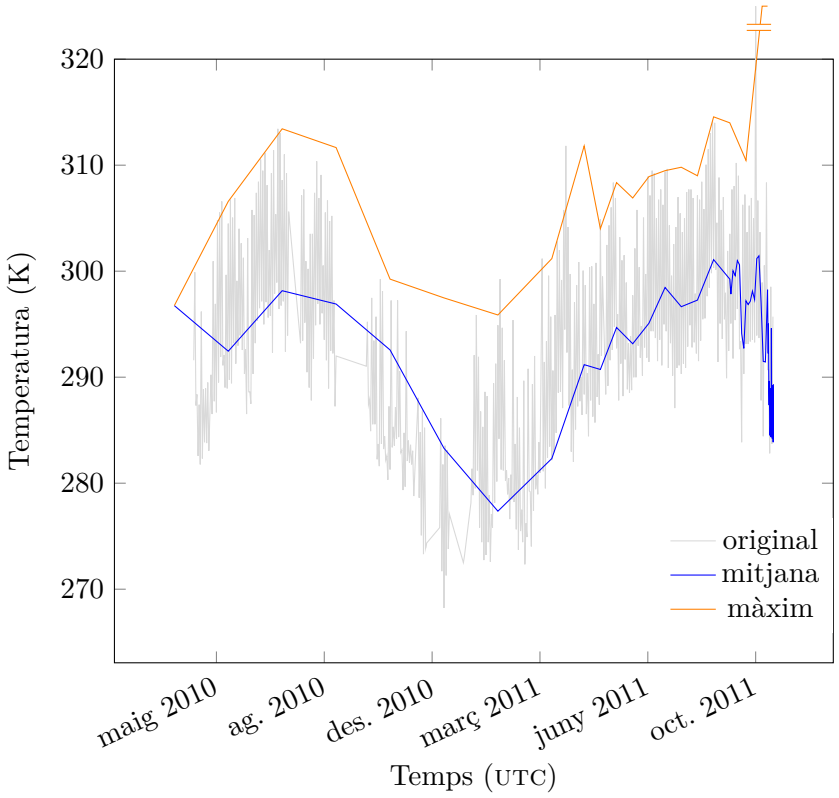


Figura 14.6.: Sèries temporals amb mètode de representació FOH

compressió de les dades originals que conté una certa informació dels atributs originals. En efecte, les 146.709 mesures emmagatzemades originalment es redueixen a 92 mesures. Aquestes mesures es reparteixen en diferents resolucions de manera que hi ha més informació per als temps recents i que cada subsèrie temporal és regular de període δ. La sèrie temporal total no és regular, però sí que se'n pot observar una regularitat a trossos ja que és una concatenació de les subsèries regulars.

14.4. Computació

Hem computat la consolidació amb dues de les implementacions de SGSTM dissenyades: RoundRobinson i RoundRobindoop. A més, en el cas de RoundRobindoop l'hem executat tant a la shell com a Hadoop.

Els resultats són aproximadament els mateixos per a totes les computacions. Hi ha, però, petites diferències degudes a les diferències entre les implementacions. Per exemple, a RounRobindoop s'ha hagut d'aproximar els intervals de consolidació, que en el cas ZOHE s'amplien un interval (v. ex. 12.2). En aquest exemple hi ha dades mancants de més d'un interval de durada i per tant cal utilitzar més d'un interval per a farcir els forats: RoundRobindoop no té aquesta possibilitat, en canvi RoundRobinson sí.

En aquest exemple ja es disposa de totes les dades originals emmagatzemades, per tant la computació és en temps diferit. Així doncs, tant podem emmagatzemar la sèrie temporal en un SGSTM com RoundRobinson i aplicar la consolidació, com calcular la funció de multiresolució de la sèrie temporal en un SGSTM com RoundRobindoop. Apliquem algunes variacions a les implementacions per tal d'adequar-les a la computació en diferit:

- Ajustament dels temps d'inici de la consolidació (v. def. 12.3). Un cop definit l'esquema de multiresolució, com que ja es coneix tota la sèrie temporal es poden canviar els temps d'inici de la consolidació. D'aquesta manera no es computen dades que immediatament seran descartades. A més a RoundRobindoop cal adequar-los per tenir en compte els cardinals màxims. Així si originalment tots els temps inicials eren $\tau_0 = 1$ de gener de 2010, ara cadascun canvia segons els paràmetres de multiresolució i segons el darrer temps de la sèrie temporal que és 18 d'octubre de 2011 a les 13:27:59. Per exemple, per a la resolució de pas $\delta_1 = 5$ h el nou temps inicial és de $\tau_1 = 13$ d'octubre de 2011 a les 10:00, el qual intuïtivament es pot veure que defineix un lapse de 5 dies segons el pas $\delta_1 = 5$ h i $\kappa_1 = 24$ mesures consolidades.
- A RoundRobindoop, com ja hem comentat, cal fer l'aproximació dels intervals de consolidació per al cas ZOHE.

Prova	RRson	RRdoop shell	RRdoop Hadoop
1	34	38	37
2	36	37	39
3	33	38	41
4	35	38	39
5	34	37	39
$\bar{x}$	34	38	39

Taula 14.1.: Proves del temps de còmput, expressat en minuts

- A RoundRobinson les subsèries resolució poden compartir el mateix buffer per a la sèrie temporal, com hem notat breument en la secció 6.1. Com que l'exemple és en temps diferit, inicialment s'emmagatzema tota la sèrie temporal en un mateix buffer. Posteriorment, en la consolidació, les subsèries resolució seleccionen d'aquest buffer les mesures adequades per a cada agregació. En aquest cas de temps diferit, es podria prescindir de l'eliminació de mesures antigues del buffer. Això no obstant, el tenim en compte de manera que el pas de consolidació més gran, el de 50 dies, és el que marca quan les mesures són antigues. A tals efectes, a RoundRobinson hem dissenyat un objecte MultiresolutionSeriesSharedBuffer, el qual és una petita variació de l'objecte MultiresolutionSeries que implementa la compartició de buffers.

Així doncs, hem computat el mateix problema de tres maneres diferents: amb RoundRobinson (RRson), amb RoundRobindoop executat a la shell (RRdoop shell) i amb RoundRobindoop executat a Hadoop (RRdoop Hadoop). A tall orientatiu, hem mesurat els temps de còmput. Per a cada cas n'hem fet cinc proves en una màquina de sobretaula, les quals es resumeixen a la taula 14.1. Per a més detall, a [83, `http://escriny.epsem.upc.edu/svn/rrb/src/experiments/tags/isense-2014-tesi/`] podeu trobar les dades i els resultats d'aquest exemple.

En conclusió, no hi ha gaires diferències entre les diferents computacions. De fet, és una computació en diferit i per tant els sistemes resolen un cas similar. En una altra estratègia, RoundRobinson permet la computació en línia en la qual el temps de còmput de la multiresolució es repartiria al llarg del temps en què s'adquireix la sèrie temporal. Pel que fa a RoundRobindoop, en el cas de l'execució a Hadoop només s'utilitza un node de computació i la quantitat de processos que escull Hadoop: 2 maps i 1 reduce. Així doncs, és una execució similar a la de la shell i la diferència pot ser vista com el cost que introdueix la gestió de Hadoop.

En una comparació més pràctica entre RoundRobinson i RoundRobindoop, aquest darrer només és un sistema de càlcul de la multiresolució. En canvi, RoundRobinson és un sistema més complet de base de dades i permet més manipulació de les sèries temporals gràcies a basar-se en Pytsms. De fet, hem dissenyat RoundRobinson i Pytsms com a les implementacions de referència dels models i conseqüentment són

més genèriques i contenen més conceptes i complements que les altres. De fet, utilitzem RoundRobinson per a crear l'esquema de multiresolució, emmagatzemar-lo en un fitxer i passar-lo com a paràmetre a RoundRobindoop. I un cop calculada la multiresolució a RoundRobindoop, tornem a utilitzar RoundRobinson per a recuperar la sèrie multiresolució consolidada resultant.

Part V.

Conclusions

15. Conclusions

En aquest capítol es resumeix l'exposat en el cos principal del document i s'extreuen conclusions del model dissenyat i de les implementacions.

Hem formalitzat la multiresolució per a les sèries temporals com un model de SGBD. La multiresolució permet seleccionar una determinada informació d'una sèrie temporal i només emmagatzemar-ne una quantitat de dades afitada. Això no obstant és una solució de compressió amb pèrdua i cal avaluar i adequar els paràmetres de la multiresolució a cada context.

El model defineix els sistemes de multiresolució de forma senzilla i compacta, la formalització permet definir un model totalment genèric i molt potent. Les implementacions de sistemes multiresolució exploren diverses maneres de computar la multiresolució, el model de multiresolució està pensat per a tenir en compte el context dels recursos: capacitat d'emmagatzematge limitada, processament en temps diferit o en línia amb el flux de dades, distribució de les dades, transmissió de resums de la informació, etc.

En comparació a altres sistemes que gestionen sèries temporals, la multiresolució és una solució de compressió que només emmagatzema les dades que es requeriran en consultes posteriors o per a visualitzacions gràfiques. El procés de descompressió és mínim ja que les dades s'emmagatzemen directament com a subsèries temporals. A més, la multiresolució col·labora a solucionar les propietats problemàtiques de les sèries temporals: la regularitat, les dades desconegudes o un volum massiu de dades. Les sèries temporals es defineixen genèriques més enllà de considerar-les simples seqüències: els instants de temps poden marcar alhora ordre i posició absoluta temporal, la distància entre les mesures pot ser irregular i consideren el seu àmbit d'adquisició en sistemes de monitoratge.

A continuació, més particularment:

- Es resumeixen els models formalitzats i la tècnica de multiresolució
- Es reflexiona sobre algunes consideracions i variacions dels models
- Es resumeix l'experimentació en el disseny d'implementacions per als models

15.1. Resum dels models

El principal objectiu de la multiresolució és l'emmagatzematge comprimit de sèries temporals. Així, la multiresolució es contextualitza en l'àmbit dels SGBD, per la qual cosa hem formalitzat el model de dades dels SGSTM, els sistemes que gestionen sèries temporals amb la tècnica de multiresolució. Els SGSTM basen el tractament de les sèries temporals en els SGST, per la qual cosa també hem descrit el model d'aquests sistemes. El principal objectiu dels SGST és gestionar les sèries temporals de manera coherent a la dimensió temporal.

La multiresolució emmagatzema una sèrie temporal mitjançant diversos resums d'atributs i resolucions. Breument, el model de SGSTM s'estructura a partir de *sèries temporals multiresolució* com a conjunt de *subsèries resolució*, les quals acumulen temporalment les mesures en un *buffer* per tal de tractar-les abans d'emmagatzemar-les a un *disc*. El tractament principal consisteix a canviar els intervals de temps entre mesures i a agregar-ne atributs, amb l'objectiu de compactar la informació de la sèrie temporal. Així, cada sèrie temporal multiresolució té diferents paràmetres per a configurar de quina manera s'ha de resumir la informació i calcular les resolucions. És el que anomenem *esquema de multiresolució* i consisteix en definir la quantitat de subsèries resolució i quatre paràmetres per a cadascuna: el *pas de consolidació*, l'*instant de temps d'inici de la consolidació*, la *funció d'agregació d'atributs* i la *capacitat d'emmagatzematge*.

El model de SGSTM també inclou les operacions que formalitzen el comportament d'aquests sistemes. En primer lloc, per a emmagatzemar una sèrie temporals multiresolució és indispensable que hi hagi operacions per a afegir mesures i per a consolidar-les. En segon lloc, hi ha operacions per a manipular l'esquema de multiresolució i per a observar-ne propietats. En tercer lloc, per a consultar les dades emmagatzemades hi ha dues operacions bàsiques: obtenir les subsèries temporals consolidades i obtenir una sèrie temporal total resultant de concatenar les subsèries. Aquestes consultes resulten en sèries temporals, així per a elaborar consultes més complexes es poden utilitzar les operacions dels SGST.

Els SGSTM necessiten les *funcions d'agregació d'atributs* per a consolidar les sèries temporals originals. Aquestes funcions, però, es formulen com a objectes independents del model per tal que l'usuari en pugui definir de pròpies. Així, hem contextualitzat les agregacions en les interpretacions de mètodes de representació de sèries temporals i hem introduït el problema de cooperar amb la validació de dades durant l'agregació. Hem exemplificat les funcions d'agregació d'atributs amb estadístics d'agregació simples –mitjana, màxim i darrer– interpretats amb tres mètodes de representació –PD, DD i ZOHE.

Pel que fa al model de SGST, s'estructura en *mesures* i *sèries temporals*, les quals tenen atributs de *temps* i *valor*. En les operacions es distingeix el comportament de les sèries temporals en conjunt, en seqüència i en funció temporal. També s'observen

les propietats de les sèries temporals en contextos determinats: els trets semàntics, els mètodes de representació i les patologies. Particularment, la multiresolució ha de cooperar per a solucionar algunes de les propietats problemàtiques de les sèries temporals, com per exemple la regularitat, un gran volum de dades o els diversos mètodes de representació.

En resum, proposem una solució de compressió que només emmagatzemi la informació prevista que es requerirà en futures consultes o bé en visualitzacions gràfiques. En comparació, generalment, altres sistemes per sèries temporals comprimeixen i reconstrueixen el senyal original o bé emmagatzemen massivament les dades. Com a tècnica de compressió, en els SGSTM el procés de descompressió és nul quan es consulta exactament una resolució de les emmagatzemades. A més, la multiresolució coopera amb propietats problemàtiques típiques de les sèries temporals i s'adequa a l'àmbit de monitoratge.

15.2. Consideracions

La multiresolució és una selecció d'una part de la informació d'una sèrie temporal. Per tant, és una solució d'emmagatzematge amb pèrdua. Com a conseqüència, l'usuari ha d'escollir un esquema de multiresolució adequat al context en què vulgui treballar.

El model presentat permet aplicar els SGSTM en dispositius que requerixen un emmagatzematge reduït i afitat. De fet, en els SGSTM es pot preveure a priori la quantitat total de mesures emmagatzemades per una durada de temps suficientment llarga, tot i que evidentment finita. Així, es pot evitar la pèrdua de la informació per manca de capacitat d'emmagatzematge o per manca de temps d'enviar les dades a un node central, cosa que per exemple pot ocorre en sistemes integrats petits.

El model és abstracte i genèric, sobretot pel que fa als buffers, per tal d'abastar el màxim de contexts. Així, principalment es defineix una consolidació independent de cada resolució en què és possible utilitzar qualsevol esquema de multiresolució.

Ara bé, considerant diversos contextos en què s'utilitzarà la multiresolució, es poden fer petits variacions en el comportament dels buffers. Aquestes variacions, però, impliquen restriccions en els paràmetres, per exemple limitem les funcions d'agregació d'atributs o els passos de consolidació que es poden utilitzar. Les variacions explorades són: diverses resolucions que comparteixen el mateix buffer; resolucions encadenades, en què unes depenen de les sèries temporals consolidades en altres; funcions d'agregació d'atributs orientades a flux, és a dir que acumulen incrementalment la computació de les agregacions; i rellotges de consolidació de diferent naturalesa –externs al sistema, interns o bé relatius a altres resolucions.

Pel que fa a la inserció de mesures i el procés posterior de consolidació, el model de SGSTM permet tant una computació en temps diferit com en línia. És a dir,

tant permet tenir una sèrie temporal ja capturada i processar-la en un SGSTM com processar-la al mateix temps que es va adquirint i per tant seguint el flux d'entrada de dades. Aquesta possibilitat de processar en línia permet fer menys crítica la computació ja que la reparteix durant l'adquisició de la sèrie temporal. A més permet disposar en línia de la multiresolució computada i per tant visualitzar-la al mateix temps en què es va adquirint.

15.2.1. Funcions de multiresolució i sistemes duals

Expressant el model de multiresolució en computació diferida, es pot observar el problema com una funció sobre una sèrie temporal que retorna una nova sèrie temporal. Així, es poden dissenyar dues funcions de multiresolució que tenen una funcionalitat equivalent a les dues consultes bàsiques dels SGSTM. Una, fa un mapa sobre la sèrie temporal original i un paràmetre de multiresolució per a obtenir la subsèrie consolidada. L'altra, fa un plec sobre la sèrie temporal original i un esquema de multiresolució per a obtenir la sèrie temporal total.

Aquestes funcions de multiresolució són operacions que es poden resoldre en els SGST. De fet, és interessant la possibilitat de computar-les amb tècniques distribuïdes i paral·leles com comentem després.

A partir de les funcions de multiresolució es poden dissenyar estructures més complexes en què intervingui la multiresolució. Aquest és el cas de sistemes duals de multiresolució en què una mateixa sèrie temporal s'emmagatzema alhora en un SGST i en un SGSTM. Aleshores, segons les necessitats, es poden consultar les sèries temporals consolidades emmagatzemades al SGSTM o bé, de forma equivalent, consultar les funcions de multiresolució al SGST. Els sistemes duals ofereixen altres aplicacions de la multiresolució.

En primer lloc, la multiresolució implica pèrdua de dades i l'SGST dels sistemes duals pot servir com a dipòsit a llarg termini de les dades originals, el qual no es consulta freqüentment. Habitualment es consultarien els resums d'informació emmagatzemats a l'SGSTM i en cas que aquests no fossin suficients es podrien usar les dades de l'SGST. Per exemple per a resoldre consultes detallades sobre les dades semblants a 'a quina hora exacta ha ocorregut un esdeveniment'.

En segon lloc, la multiresolució requereix definir un esquema de multiresolució abans de començar a recollir les dades. Però cal conèixer l'entorn monitorat per a establir-ne un esquema de multiresolució, per exemple per a determinar quines funcions d'agregació són les més escaients. En aquest període de transició, fins que es conegui l'esquema de multiresolució més escaient, es pot utilitzar un sistema dual per a experimentar amb diferents esquemes. I un cop determinats, ja es pot utilitzar només l'emmagatzematge en l'SGSTM.

En tercer lloc, la computació en flux dels SGSTM és una estratègia que convé tenir present ja que permet repartir el temps de computació en línia amb l'adquisició de les dades. Així, prenent com a referència altres conceptes similars en les vistes dels SGBD i en l'arquitectura Lambda, els sistemes duals es poden aplicar seguint l'estratègia següent. L'SGSTM processa el flux de dades al mateix temps en què s'adquireixen i sempre ofereix la resposta de multiresolució forma immediata, és a dir té els resultats precomputats. L'SGST actua per acabar de completar les consultes o bé si s'ha de canviar l'esquema de multiresolució i tornar a començar de nou.

Això no obstant, cal tenir present que els SGST són un emmagatzematge massiu de les dades. A més, els sistemes duals impliquen en certa manera un emmagatzematge redundant de la informació en els dos sistemes. En els sistemes duals, assumim que l'usuari vol obtenir la multiresolució de la sèrie temporal original: ja sigui per a visualitzar-la directament o per a utilitzar-la com a base per a altre consultes. Així, hi ha contextos en què els resum de la multiresolució són suficients i es pot prescindir de conservar totes les dades històriques, per exemple en el monitoratge de l'evolució de la bateria disponible en un portàtil és suficient visualitzar-ne un resum. També en altres contextos la multiresolució és suficient per a respondre determinades consultes sobre les dades, per exemple una consulta similar a 'en aquestes dades la mitjana dels darrers dies és similar a l'habitual o bé creix'. En aquests contextos, el model de multiresolució és molt adequat; la reflexió sobre la qualitat de la multiresolució aclareix quan les dades emmagatzemades són suficients i quan la informació emmagatzemada és redundant.

15.2.2. Reflexió sobre la qualitat

La multiresolució és una tècnica de compressió amb pèrdua de dades i els paràmetres de l'esquema de multiresolució són graus de llibertat per a cada context d'aplicació. Així, depenent d'aquests paràmetres, se selecciona una informació o una altra de la sèrie temporal original. La compressió amb pèrdua implica que algunes operacions de SGST són consultes aproximades quan es resolen a partir de les sèries temporals emmagatzemades en els SGSTM. Com a conseqüència, cal determinar la qualitat que té la compressió de la multiresolució per tal de poder dissenyar correctament els esquemes de multiresolució i poder quantificar l'aproximació de les consultes que s'hi basin.

En altres àmbits, per al cas de pèrdua principalment la compressió de multimèdia, la teoria de la informació avalua formalment l'efecte que té la compressió en la informació que hi ha originalment a les dades. L'anàlisi que formulem, però, és una introducció a la reflexió sobre l'error en la informació de la multiresolució. Així, de forma simple, analitzem si hi ha error o si no n'hi ha, sense pretendre quantificar-lo més detalladament.

Si bé definim el problema genèric d'error en la multiresolució, és un problema massa abstracte per a treballar-lo directament. Així doncs, analitzem alguns casos particulars per tal que serveixin com a exemple per a reflexionar sobre l'efecte que té una configuració determinada de paràmetres multiresolució. Concretament, es pot avaluar l'error de multiresolució per a algunes funcions d'agregació d'atributs particulars i de com siguin les consultes posteriors. Cal destacar el cas específic dels comptadors, en què es pot conservar la informació de comptatge que contenen coherentment amb la seva naturalesa.

15.3. Experimentació

Un cop s'han definit els models dels SGST i dels SGSTM, es poden implementar. D'aquesta manera, els models descriuen exactament el comportament de les implementacions i es pot assegurar que aquestes tenen el funcionament desitjat. De fet, els models definits descriuen essencialment les operacions, entrades, sortides i tipus de dades bàsics amb els quals han de treballar les implementacions. Això és conegut com a interfície de programació d'aplicacions, tot i que habitualment s'especifiquen i es representen amb llenguatges informàtics, en comptes de models formals matemàtics, i incorporen detalls de la implementació en qüestió.

L'objectiu de les implementacions és principalment acadèmic per a mostrar el funcionament dels models, tot i que també són útils per poder experimentar amb dades reals. Així, dissenyem tres implementacions de naturalesa diferent.

En primer lloc, Pytsms i RoundRobinson que són la implementació de referència, és a dir la que considerem que mostra com es pot implementar el model de multiresolució definit. Tant admet la computació en línia com en diferit. S'ha implementat amb llenguatge Python, tot aprofitant el paradigma d'orientació a objectes per a obtenir una bona correspondència entre implementació i model. A més Python té unes biblioteques extenses que permet afegir complements com per exemple gràfics o gestionar l'emmagatzematge en fitxers.

En segon lloc, RoundRobindoop implementat amb MapReduce i Hadoop. Utilitza MapReduce com a tècnica de programació paral·lela i distribuïda i usa Hadoop com a sistema de computació d'aquesta tècnica. Està basat en les funcions de multiresolució i per tant és una computació en diferit. Hi ha, però, algunes simplificacions del model de multiresolució per a poder-se ajustar a MapReduce.

En tercer lloc, Reltsms implementat amb Tutorial D, el llenguatge acadèmic dels SGBDR, i Rel com a intèrpret. Aquesta és la implementació més experimental, de fet Rel està en desenvolupament. Permet situar els SGST en els SGBDR i alhora avaluar l'adequació del model relacional per a tipus complexos com són les sèries temporals i la multiresolució. Pel que fa a la multiresolució, s'han implementat les

funcions de multiresolució. Aquesta implementació, però, és la més acadèmica i és limitada a l'hora de provar amb dades reals.

Finalment, posem a prova les implementacions amb dades reals massives. Aquestes proves demostren el bon funcionament de la multiresolució en un context determinat. A més, permeten exemplificar la cobertura del cronograma de multiresolució sobre unes dades originals i mostrar les característiques dels diferents mètodes de representació i funcions d'agregació d'atributs, particularment del màxim i la mitjana de la família ZOHE.

En una comparació entre les diverses implementacions de la multiresolució, s'ha observat que computen els mateixos resultats per a un mateix cas. Tot i així, Reltsms no admet casos reals massius, ja que es basa en un SGBDR totalment acadèmic i encara en desenvolupament. Cal notar, però, que en la programació de les implementacions no hem tingut en compte l'eficiència, de fet hem preferit conservar una bona similitud amb les operacions algebraiques definides en els models. Així doncs, basar-se en la teoria dels SGBD a l'hora de definir els models ha resultat adequat per a obtenir unes implementacions abstractes, sense gaires detalls físics. A més, ha permès implementacions de diversa natura, com són Pytsms i Reltsms. RoundRobindoop és la implementació més específica de totes, la podem qualificar com a sistema de càlcul en paral·lel de la multiresolució. En canvi, RoundRobinson i Pytsms són les més genèriques, car les hem dissenyat com a referència dels models. Però també gràcies que Python disposa de biblioteques extenses, les quals són útils a l'hora d'implementar funcionalitats complementàries.

Pel que fa a la computació de la multiresolució, s'ha vist que pot ser costosa a causa dels múltiples càlculs que s'han de computar. Hem estudiat dues estratègies remarcables. D'una banda, la computació paral·lela i distribuïda, que distribueix el còmput en els recursos físics disponibles i per tant permet processar ràpidament. D'altra banda, la computació en línia amb el flux de dades, que distribueix el còmput al llarg del temps d'adquisició i per tant el fa menys crític. En aquest cas, la multiresolució és precomputada i sempre està disponible per a consultar-la o visualitzar-la immediatament. Ara bé, requereix planificar prèviament l'esquema de multiresolució.

16. Treball futur

El treball presentat obre la possibilitat a nous temes de recerca, alguns dels quals creiem que són reptes interessants. En aquest capítol es reflexiona sobre possibles treballs de recerca futurs. Proposem nous treballs futurs al voltant de tres temes: els models, les implementacions i reflexions sobre la qualitat de la multiresolució.

16.1. Models

En el model de SGST s'ha definit un seguit d'operacions que són les que considerem més bàsiques per a poder manipular les sèries temporals. Algunes operacions són conseqüència directa d'altres conceptes; per exemple la diferència a partir de la pertinença o la intersecció a partir de la diferència. D'altres operacions també en són conseqüència però cal prendre una determinada decisió sobre el raonament; per exemple en la unió i la unió temporal cal decidir quina prové de l'ordre total i quina de l'ordre parcial. I d'altres, funcionen com a passos previs per a altres operacions; per exemple la concatenació temporal s'utilitza en els SGSTM per a calcular la sèrie temporal total en el context d'un mètode de representació. Així doncs, caldria establir clarament la motivació de cada operació i cercar en cada cas el raonament sobre el qual es pot basar cada operador. També, de forma breu, hem notat les propietats d'alguns operadors com la commutativitat, però caldria explorar més profundament les propietats de tots els operadors.

En les funcions d'agregació d'atributs dels SGSTM, n'hem proposat alguns exemples sobretot raonats a partir dels mètodes de representació. Tot i així, hem proposat estadístics senzills –mitjana, màxim i darrer valor– atès que l'objectiu és mostrar el comportament que tenen en la multiresolució. En els treballs d'anàlisi de sèries temporals es presenten multitud de mètodes i d'algoritmes: per a extreure patrons de les sèries temporals, per a cercar periodicitats, per a comparar dues sèries temporals, agregacions en el domini freqüencial, per a fer prediccions, per a validació de dades, etc. Per tant, es poden dissenyar més funcions d'agregació d'atributs basant-se en qualsevol d'aquests algoritmes o mètodes; només cal adaptar el problema per tal de retornar una mesura que resumeixi la informació d'un interval de la sèrie temporal.

De les funcions d'agregació d'atributs n'hem notat la possibilitat d'orientar-les a flux. El model de SGSTM és adequat per a computar-se en flux llevat de les funcions

d'agregació d'atributs que es defineixen genèriques. Creiem, doncs, que també és interessant aplicar orientació a flux en aquestes funcions i caldria aprofundir en aquests algoritmes, com per exemple els que proposa Cormode, Korn i Tirthapura [18].

En la teoria de la mesura, la incertesa sol acompanyar les mesures. La incertesa reflecteix probabilísticament els límits del que es coneix sobre la quantitat mesurada. Així doncs, seria interessant poder incorporar la incertesa en els models. Principalment la incertesa hauria d'acompanyar els atributs de temps i de valor de les sèries temporals, és a dir haurien de reflectir la incertesa que hi ha en cada mesura a l'hora d'adquirir un valor un instant de temps. Aleshores, caldria estudiar com aquesta incertesa afecta les operacions, és a dir com es propaga la incertesa quan s'uneixen dues sèries temporals, quan es representen, en les funcions d'agregació d'atributs, etc.

Pel que fa a la consolidació de la multiresolució, aquesta s'ha pensat sobretot periòdica per tal d'obtenir sèries temporals regulars. Aleshores s'obtenen els esquemes de multiresolució periòdics que hem analitzat. Això no obstant, són possibles altres escenaris de consolidació; un exemple seria el cas sistemes de monitoratge que adquireixen dades només quan ho creuen interessant en base a esdeveniments. En el model de SGSTM no estan previstos aquests casos i requereixen un estudi més profund. Per exemple, en un cert moment es podria considerar que una subsèrie resolució és molt significativa i que s'han de mantenir aquestes dades, per tant en l'esquema de multiresolució hi hauria una subsèrie fixada a un interval de temps en el passat.

En els esquemes de multiresolució, s'ha notat la propietat de desfasament d'una subsèrie resolució. Aquest desfasament és induït per la funció d'agregació d'atributs. D'una banda, aquest desfasament pot ser requerit per la naturalesa de la funció; per exemple en el cas d'agregacions DD o ZOH introdueixen un desfasament perquè interpolen cap endavant, de fet el resultat de consolidació entre la ZOH i la ZOHE és similar però aquesta darrera per naturalesa interpola cap enrere i no necessita desfasament. D'altra banda, aquest desfasament podria ser afegit intencionalment per a controlar els solapament entre les subsèries resolució. Aquest solapament significa que en l'esquema de multiresolució les diverses subsèries resolució coincideixen a emmagatzemar informació pels mateixos instants de temps. Proposem dos escenaris de solapament:

- Les subsèries resolució se solapen totalment. Això pot servir per a disposar de diferents resums de la sèrie temporal preparats per a ser visualitzats immediatament. És a dir, el SGSTM permet escollir ràpidament entre diferents *zooms* de les dades.
- Les subsèries resolució no se solapen, és a dir les subsèries amb menys resolució acaben allà on comencen les de més resolució. Això pot servir per a aprofitar al màxim la resolució i l'espai d'emmagatzematge, sense que cap subsèrie desi

informació per al mateix interval de temps. A part d'introduir desfasament, per tal que no se solapin també es pot dissenyar un esquema de multiresolució on el pas de consolidació de cada subsèrie sigui exactament $\delta_j = \kappa_i \delta_i$ on $\delta_i < \delta_j$. És a dir la subsèrie amb menys resolució (j) té un pas de consolidació exactament múltiple del pas de consolidació i el cardinal de la subsèrie superior en resolució (i). Aquestes restriccions es podrien considerar també en el cas de l'estructura de resolucions encadenades.

16.2. Implementacions

En les implementacions hem treballat a nivell acadèmic, és a dir sense objectius d'optimització del rendiment. De fet, aleshores, les implementacions s'allunyarien del model i per tant del nostre objectiu de mantenir una forta correspondència entre la forma de la implementació i del model. Aquesta correspondència és útil per a manteniments futurs: qualsevol millora en el model pot ser traslladada immediatament a les implementacions o bé, a la inversa, qualsevol error trobat en les implementacions pot ser localitzat fàcilment i estudiat en el model.

Ara bé, aquesta correspondència model-implementació no sempre és senzilla de mantenir. A Pytsms i RoundRobinson, les quals són les implementacions més completes, s'hauria de simplificar la definició de les mesures genèriques. En el model abstracte matemàtic és senzill de descriure uns valors genèrics, en canvi a les implementacions això complica l'estructura. Així, s'ha hagut de dissenyar mesures de diferents tipus que contenen el rang del domini de temps i valors per tal de definir les mesures indefinides, el suprem i l'ínfim, etc. També s'hauria de repensar la gestió de la homogeneïtat de les sèries temporals: en el model les sèries temporals són homogènies i en les implementacions és difícil gestionar el concepte de nova sèrie temporal amb el mateix tipus de mesures que les originals.

En altres casos, però, s'ha trobat una solució adequada per a implementar la genericitat del model. És el cas de la multitud d'operacions de les sèries temporals implementades amb *Mixin*, el de les representacions com a objectes independents associats a les sèries temporals, o bé els de les funcionalitats complementàries com l'emmagatzematge i els gràfics implementades amb el patró *Visitor*. Tanmateix, algunes parts encara no han estat prou generalitzades; per exemple els gràfics de RoundRobinson sempre utilitzen la representació ZOHE.

En les altres implementacions d'SGSTM, l'objectiu s'ha centrat en observar altres paradigmes d'implementació dels models. Així, s'ha optat per no implementar tota la genericitat del model sinó casos simplificats. Caldria avaluar fins a quin límit aquestes implementacions es podrien apropar més al model, per exemple a RoundRobindoop hem notat algunes limitacions a l'hora d'usar les funcions d'agregació d'atributs.

A RoundRobindoop usem Hadoop com a intèrpret de la tècnica de programació paral·lela MapReduce. L'execució d'aquesta tècnica implica un compromís a l'hora d'escollir el nombre de processos en paral·lel ja que cada un té un cost mínim de crear-se i a més cal comptar el cost de distribuir les dades. Es podria experimentar més amb Hadoop en aquest sentit, és a dir amb diferents quantitats de *maps* i de *reduces*. De fet només hem provat amb un node de computació, però Hadoop té la possibilitat de distribuir a més nodes.

Hi ha altres projectes que també utilitzen Hadoop com a sistema d'emmagatzematge distribuït de sèries temporals. És el cas d'OpenTSDB [127], que utilitza Hadoop per a emmagatzemar i recuperar ràpidament sèries temporals. Aquest, però, tan sols té en compte recuperar les dades originals emmagatzemades i no preveu consultes elaborades. RoundRobindoop és una solució per a computar la multiresolució a Hadoop. Així doncs, OpenTSDB i RounbdRobindoop podrien treballar conjuntament: el primer per a emmagatzemar distribuïdament les sèries temporals i el segon per a calcular la multiresolució aprofitant que les dades ja estan distribuïdes en diversos nodes.

A Reltsms hem implementat el model d'SGST seguint la programació acadèmica del model relacional. Es podria seguir la mateixa aproximació per a implementar també el model d'SGSTM. En aquestes implementacions, es podria experimentar amb un dels punts forts dels SGBDR: l'optimització de les consultes [23, cap. 18 *Optimization*]. Les expressions relacionals són d'alt nivell matemàtic i això permet trobar expressions equivalents a una consulta. Aleshores, els sistemes poden decidir quina expressió és la millor per a ser executada. En aquest sentit, hem definit els operador de Reltsms a partir dels operadors relacionals. Això no obstant, s'hauria d'estudiar si a Tutorial D les funcionalitats d'optimització s'estenen automàticament als operadors derivats dels primitius.

En un sentit relacional, també cal comparar Pytsms i RoundRobinson amb Dee. Dee [52] és la implementació amb Python d'un llenguatge de bases de dades relacionals que compleix amb les normes D del Third Manifesto. A Pytsms i RoundRobinson hem utilitzat els conjunts de Python com a objectes bàsics però es podrien utilitzar els conjunts relacionals de Dee, els quals ja incorporen propietats i mètodes de SGBD. Si més no, el raonament que fa Dee com a gestor de bases de dades també podria ser aplicat a Pytsms i RoundRobinson. Tot i així, actualment Dee sembla un projecte abandonat però es poden trobar projectes recents amb raonaments similars com per exemple Alf [75].

Amb l'experimentació amb dades reals hem pogut demostrar el correcte funcionament de la multiresolució. Hem provat les implementacions amb un exemple de dades reals massives però caldria experimentar amb més diversitat de dades: dades de diferent mida, de diversa naturalesa, diferents esquemes de multiresolució, etc. També cal tenir present que en l'àmbit d'anàlisi de sèries temporals hi ha dades de referència [64] per a ser utilitzades en altres recerques juntament amb el problema concret que cal resoldre-hi, per exemple cerques de similituds.

Amb més diversitat de dades també es podria avaluar el rendiment dels recursos durant la computació de la multiresolució. Com ja hem comentat, el nostre objectiu se centra en propòsits acadèmics i deixa de banda qüestions més productives. Ara bé, durant la programació de les implementacions hem pogut constatar alguns efectes que caldria estudiar amb més detall. En el casos de Python, la implementació ens ha resultat més eficient com més bé hem pogut traslladar les definicions a recursos d'alt nivell de Python. Aquest és el cas d'implementar les operacions de mapa i plec en les funcions de Python equivalents *map* i *reduce*. També hem notat una gran millora canviant la inserció de les *Measure* a les *TimeSeries* per tal que en la cerca que no hi hagi temps repetits s'aprofiti la cerca en els sets de Python mitjançant hash. En el cas de MapReduce, hem seguit utilitzant RoundRobinson per als disseny de les funcions d'agregació d'atributs. D'aquesta manera, mantenim la possibilitat de funcions genèriques, tot i que caldria avaluar si Hadoop obtindria més profit d'altres implementacions. Per exemple, algoritmes de resolució de funcions d'agregació d'atributs que aprofitessin el flux de les operacions d'ordenació o bé que en l'etapa de map ja realitzessin una part dels càlculs.

Pel que fa a les variacions dels SGSTM estudiades, n'hem implementat algunes a RoundRobinson: l'ajustament dels temps d'inici i la compartició d'un mateix buffer per a emmagatzemar la sèrie temporal original. Les altres variacions, com les resolucions encadenades o l'orientació a flux de les funcions d'agregació d'atributs, també podrien ser implementades. Aleshores, seria interessant avaluar la computació de la multiresolució per a aquests diferents casos. És a dir, realitzar diferents proves amb dades de diversa naturalesa i comparar-ne el funcionament. Un cas semblant és el de les funcions de multiresolució, les quals hem implementat a RoundRobindoop i a Reltsms però també es podrien implementar a Pytsms. Aleshores, es podria observar clarament en una mateixa implementació l'equivalència de funcionalitat entre el model de SGSTM i les funcions de multiresolució De fet, aquesta equivalència s'hauria d'estudiar més profundament en la formalització dels models.

16.2.1. Sistemes de multiresolució integrats en maquinari

Es poden realitzar altres implementacions dels SGSTM que siguin molt específiques. Una bona implementació ja no és només aquella que calcula en poc temps sinó que en alguns contextos també pot ser un consum baix d'energia, ocupar poc espai, etc. Més concretament, ens referim a sistemes específics integrats en xarxes de sensor.

En aquest sentit, es podria implementar un SGSTM integrat en el maquinari d'un sensor. Aquesta implementació integrada es podria realitzar en un microcontrolador que gestionés una memòria seguint l'esquema de multiresolució. O bé, es podria aprofitar que l'esquema de multiresolució té una mida finita i és implementable en maquinari, per a dissenyar-la com a circuit digital.

En la implementació en maquinari, es podria seguir l'esquema de la figura 16.1. Aquest esquema és per a una subsèrie resolució, per tant una sèrie temporal multiresolució seria un conjunt d'aquests esquemes. Així, aquest esquema és la integració de l'esquema de les subsèries resolució de la figura 5.3. En el buffer es van afegint les mesures –temps i valor– i calcula la mesura resultant –l'atribut agregat. Aquest atribut agregat s'emmagatzema al disc a cada pas de consolidació –marcat per l'esdeveniment consolida– i el disc gestiona l'emmagatzematge afitat –les dades consolidades $D_0, D_1, \ldots, D_k$. Caldria afegir un mòdul de temps que a partir del rellotge – per exemple un real-time clock (RTC)– marqués els passos de consolidació i indiqués el darrer temps de consolidació.

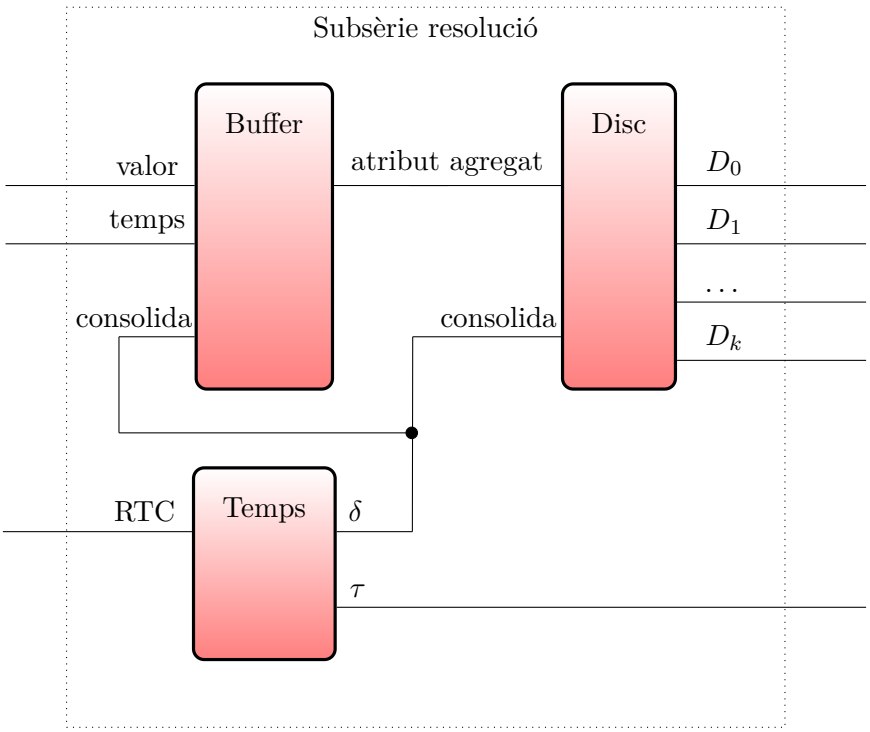


Figura 16.1.: Esquema d'integració d'una subsèrie resolució

En aquesta implementació només fem referència a la part d'emmagatzematge. Caldria implementar un protocol per tal de consultar les dades emmagatzemades, si bé forma senzilla es podria implementar com si els discs fossin un perifèric de memòria.

Algunes aplicacions dels sistemes integrats de multiresolució podrien ser:

- Emmagatzematge de la multiresolució en perifèrics la informació dels quals, tot i no ser essencial, pot ajudar a monitorar-ne el seu funcionament. Per exemple comptadors d'aparells de xarxa, temperatures dels components, etc.

- Aparells integrats molt petits, en els quals hi ha molt poc espai per a l'emmagatzematge.
- Com un complement més de sensors intel·ligents, que actualment ja integren diverses tasques: filtratge del senyal, busos de comunicacions, llindars d'alarma, etc.
- Per a computar funcions d'agregació d'atributs complexes. En aquest cas els buffers podrien treballar directament amb components del maquinari. Per exemple per a agregacions de sèries temporals en què els valors fossin imatges.
- Implementació de la multiresolució en Field Programmable Gate Arrays, és a dir en dispositius de maquinari configurables. Això permetria major flexibilitat a l'hora de canviar els esquemes de multiresolució integrats.

16.3. Reflexions sobre la qualitat

El capítol d'aplicació de la teoria de la informació és una introducció al problema de la qualitat de la multiresolució. La teoria de la informació formalitza anàlisis més profundes per a la compressió de dades que es podrien aplicar també a la multiresolució. En el cas que es conegui més bé el context i el comportament de la sèrie temporal a la qual s'aplica la multiresolució, es pot detallar més bé la quantificació de l'error. Aleshores, en termes de la teoria de la informació, hi ha més coneixement sobre la predicció del comportament de les dades cosa que es pot utilitzar per a avaluar característiques més concretes. Per exemple, una variable real adquirida té limitat el rang de valors que pot prendre i fins i tot pot tenir un comportament probabilístic determinat.

Cal notar que no avaluem la idoneïtat d'aplicar un estadístic o un altre a unes dades. Altrament, ja partim del cas que es vol aplicar una consulta amb una agregació determinada a una sèrie temporal. Per a determinar un esquema de multiresolució –la quantitat de resolucions, els passos de consolidació de cadascuna, els cardinals, ...– caldria analitzar cada problema particular en el seu context i utilitzar els coneixements adequats. Per exemple, per a treballar amb problemes de so la teoria del senyal formalitza tot de raonaments que no es poden obviar a l'hora de definirne un esquema de multiresolució. O bé, també en la teoria del senyal, es poden trobar anàlisis per a determinar bons passos de consolidació per a una variable, per exemple de temperatura. Així i tot, en el cas dels comptadors se'n pot fer un raonament a banda per a conservar la seva informació genuïna de comptatge lligada a com l'adquireixen.

A partir de les reflexions fetes sobre la qualitat de la multiresolució s'obren un seguit de qüestions més, cadascuna de les quals és un repte futur:

- Quina redundància d'emmagatzematge hi ha entre diverses subsèries d'una mateixa sèrie temporal multiresolució? Què ocorre quan hi ha més d'una resolució i són de diferents funcions d'agregació d'atributs?
- En cas que es perdi una resolució, es podria reconstruir a partir de les altres? O bé en cas que es vulgui ampliar la mida d'un disc, es podria completar amb les dades d'altres resolucions?
- Les resolucions encadenades són interessant perquè aprofiten les dades emmagatzemades però afegeixen més restriccions a la compressió d'informació. Com es poden barrejar diferents passos de consolidació i diferents funcions d'agregació d'atributs?
- Gràcies a un esquema de multiresolució es pot emmagatzemar dades d'una sèrie temporals durant un llarg temps de forma comprimida. Aquesta durada de temps és llarga però finita, tot i que quan s'esgoti dinàmicament es pot afegir una nova subsèrie amb inferior resolució però amb més lapse. Inicialment aquesta subsèrie serà buida, però es podria utilitzar les dades ja emmagatzemades per a iniciar-la?

Bibliografia

[1] Cesare Alippi et al. "An hybrid wireless-wired monitoring system for real-time rock collapse forecasting". A: *7th International Conference on Mobile Adhoc and Sensor Systems*. MASS '10. IEEE, nov. de 2010, pp. 224 - 231. DOI: 10.1109/MASS.2010.5663999 (citat a la p. 227).

[2] Steve Allen. *Time Scales*. Ver. 1.357. University of California Observatories. 4 de set. de 2013. URL: http://www.ucolick.org/~sla/leapsecs/timescales.html (cons. 10-1-2014) (citat a la p. 58).

[3] Johannes Aßfalg. "Advanced Analysis on Temporal Data". Tesi doct. Fakultät für Mathematik, Informatik und Statistik der Ludwig Maximilians Universität München, 19 de mai. de 2008. URL: http://edoc.ub.uni-muenchen.de/8798/ (cons. 14-4-2011) (citat a les pp. 16, 21, 24, 39, 61).

[4] Johannes Aßfalg et al. "T-Time: Threshold-Based Data Mining on Time Series". A: *Proceedings of the 24th International Conference on Data Engineering*. ICDE '08. Cancun, Mexico: IEEE, abr. de 2008, pp. 1620 - 1623 (citat a la p. 39).

[5] Paolo Atzeni et al. "The relational model is dead, SQL is dead, and I don't feel so good myself". A: *SIGMOD Record* 42.1 (mar. de 2013), pp. 64 - 68. ISSN: 0163-5808. DOI: 10.1145/2503792.2503808 (citat a les pp. 16, 27, 30, 36).

[6] Brian Babcock et al. "Models and issues in data stream systems". A: *Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*. PODS '02. Madison, Wisconsin: ACM, 2002, pp. 1 - 16. DOI: 10.1145/543613.543615. URL: http://infolab.usc.edu/csci599/Fall2002/paper/DML2_streams-issues.pdf (cons. 17-4-2012) (citat a la p. 28).

[7] Yijian Bai et al. "Efficient support for time series queries in data stream management systems". A: *Stream Data Management*. Ed. de Nauman Chaudhry, Kevin Shaw i Mahdi Abdelguerfi. Vol. 30. The Kluwer International Series on Advances in Database Systems. Kluwer Academic Publishers, 2005. Cap. 6. DOI: 10.1007/b106968. URL: http://www.cs.ucla.edu/~zaniolo/papers/eslTS.pdf (cons. 19-4-2012) (citat a la p. 28).

[8] Alex van den Bogaerdt. *RRDtool, Tutorials and explanations*. URL: http://www.vandenbogaerdt.nl/rrdtool/ (cons. 14-11-2014) (citat a la p. 171).

[9] Philippe Bonnet, Johannes Gehrke i Praveen Seshadri. "Towards Sensor Database Systems". A: *Proceedings of the Second International Conference on Mobile Data Management*. MDM '01. Hong Kong: Springer-Verlag, gen. de 2001, pp. 3 - 14. DOI: 10.1007/3-540-44498-X_1. URL: http://www.cs.cornell.edu/johannes/papers/2001/MDM2001-sensor.pdf (cons. 16-4-2012) (citat a les pp. 25, 28, 42).

[10] Brown University et al. *H-Store: Next Generation OLTP Database Research*. 2014–2010. URL: http://hstore.cs.brown.edu/ (cons. 12-5-2014) (citat a la p. 37).

[11] Jake D. Brutlag. "Aberrant Behavior Detection in Time Series for Network Monitoring". A: *Proceedings of the 14th Systems Administration Conference*. LISA '00. New Orleans, Los Angeles: USENIX Association, des. de 2000, pp. 139 - 146. URL: http://www.usenix.org/events/lisa00/brutlag.html (citat a les pp. 40, 138).

[12] Alessandro Camerra et al. "iSAX 2.0: Indexing and Mining One Billion Time Series". A: *Proceedings of the 10th IEEE International Conference on Data Mining*. ICDM '10. Sydney, Australia: IEEE, des. de 2010, pp. 58 - 67. URL: http://www.cs.ucr.edu/~eamonn/iSAX_2.0.pdf (cons. 15-3-2011) (citat a les pp. 28, 40).

[13] David W. Cantrell. *Affinely Extended Real Numbers*. MathWorld – A Wolfram Web Resource, created by Eric W. Weisstein. 2012. URL: http://mathworld.wolfram.com/AffinelyExtendedRealNumbers.html (cons. 6-2-2014) (citat a les pp. 58, 71).

[14] David W. Cantrell. *Projectively Extended Real Numbers*. MathWorld–A Wolfram Web Resource, created by Eric W. Weisstein. 2012. URL: http://mathworld.wolfram.com/ProjectivelyExtendedRealNumbers.html (cons. 6-2-2014) (citat a la p. 60).

[15] Dale Carder. *RRDtool Scalability, performance in the large scale*. URL: http://net.doit.wisc.edu/~dwcarder/rrdcache/ (cons. 22-3-2011) (citat a la p. 41).

[16] Edgar Frank Codd. "A relational model of data for large shared data banks". A: *Communications of the ACM* 13.6 (jun. de 1970). [Reprinted on NoCOUG Journal http://www.noucoug.org/Journal/NoCOUG_Journal_201111.pdf], pp. 377 - 387. DOI: 10.1145/362384.362685 (citat a la p. 30).

[17] Edgar Frank Codd. *Derivability, Redundancy, and Consistency of Relations Stored in Large Data Banks*. RJ599. [Reprinted on ACM SIGMOD Record Vol.38 2009]. San Jose, US-CA: IBM Research, 19 d'ago. de 1969. DOI: 10.1145/1558334.1558336 (citat a la p. 30).

[18] Graham Cormode, Flip Korn i Srikanta Tirthapura. "Time-Decaying Aggregates in Out-of-order Streams". A: *Proceedings of the 27th ACM Symposium on Principles of Database Systems*. PODS '08. Vancouver, Canada: ACM, 2008, pp. 89 - 98. DOI: 10.1145/1376916.1376930. URL: http://dimacs.rutgers.edu/~graham/pubs/papers/decaypods.pdf (cons. 4-6-2014) (citat a les pp. 42, 151, 250).

[19] S. Chilingaryan et al. "Advanced data extraction infrastructure: Web based system for management of time series data". A: *Journal of Physics: Conference Series. 17th International Conference on Computing in High Energy and Nuclear Physics (CHEP '09)* 219.4 (2010). DOI: 10.1088/1742-6596/219/4/042034 (citat a la p. 40).

[20] Hugh Darwen. *Extending Tutorial D to Support User-Defined Generic Relation and Tuple Operators*. www.thethirdmanifesto.com, 18 de nov. de 2013. 4 pp. URL: http://www.dcs.warwick.ac.uk/~hugh/TTM/User-defined-relational-operators-in-TD.pdf (cons. 7-5-2014) (citat a les pp. 33, 223).

[21] Christopher J. Date. "A critique of Claude Rubinson's paper nulls, three - valued logic, and ambiguity in SQL: critiquing Date's critique". A: *SIGMOD Record* 37.3 (set. de 2008), pp. 20 - 22. DOI: 10.1145/1462571.1462574 (citat a la p. 35).

[22] Christopher J. Date. *An Introduction to Database Systems*. 7a ed. Boston, US-MA: Addisson-Wesley, 2000. ISBN: 0-201-38590-2 (citat a les pp. 28, 29).

[23] Christopher J. Date. *An Introduction to Database Systems*. 8a ed. New York, US: Pearson/Addisson-Wesley, 2004. ISBN: 0-321-18956-6 (citat a les pp. 16, 30, 32, 36, 47, 67, 68, 76 - 78, 163, 164, 199, 213, 252).

[24] Christopher J. Date. *Date on Databases: Writings 2000-2006*. 1a ed. Berkely, US-CA: Apress, 2006. ISBN: 159059746X (citat a les pp. 30, 32 - 36).

[25] Christopher J. Date. "Foundation matters". A: *Proceedings of the 28th international conference on Very Large Data Bases*. VLDB '02. Hong Kong, China: VLDB Endowment, 2002, p. 2. URL: http://dl.acm.org/citation.cfm?id=1287369.1287371 (citat a la p. 33).

[26] Christopher J. Date. *The Relational Database Dictionary, Extended Edition*. 1a ed. Berkely, US-CA: Apress, 2008. ISBN: 1430210419, 9781430210412 (citat a les pp. 30, 35).

[27] Christopher J. Date. *View Updating and Relational Theory*. 1a ed. Sebastopol, US-CA: O'Reilly, 2013. ISBN: 978-1-449-35784-9 (citat a la p. 119).

[28] Christopher J. Date i Hugh Darwen. "Databases, Types and the Relational Model". A: *Databases, Types and the Relational Model*. 3a ed. Pearson Education, 2006. Cap. 4,5,A,I. ISBN: 0321399420. URL: http://www.thethirdmanifesto.com/ (cons. 1-4-2012) (citat a les pp. 32, 35).

[29] Christopher J. Date i Hugh Darwen. "No! to SQL! No! to NoSQL!" A: *NoCOUG Journal* 27.3 (ago. de 2013), pp. 4 - 12. URL: http://www.nocoug.org/Journal/NoCOUG_Journal_201308.pdf (cons. 9-5-2014) (citat a les pp. 35, 36).

[30] Christopher J. Date i Hugh Darwen. *The Third Manifesto*. Ver. 2. www.thethirdmanifesto.com, 7 de feb. de 2013. URL: http://www.dcs.warwick.ac.uk/~hugh/TTM/TTM-2013-02-07.pdf (cons. 23-6-2014) (citat a les pp. 33, 35, 213).

[31] Christopher J. Date i Hugh Darwen. *The Third Manifesto and Tutorial D (reference material)*. 2012–2009. URL: http://www.thethirdmanifesto.com/ (cons. 1-4-2012) (citat a la p. 35).

[32] Christopher J. Date i Hugh Darwen. "Tutorial D". A: *Database Explorations: Essays on The Third Manifesto and related topics*. Trafford publishing, 2010. Cap. 11. ISBN: 978-1-42693-723-1. URL: http://www.dcs.warwick.ac.uk/~hugh/TTM/DBE-Chapter11.pdf (cons. 1-4-2012) (citat a les pp. 35, 213).

[33] Christopher J. Date, Hugh Darwen i Nikos A. Lorentzos. *Temporal Data and the Relational Model. A detailed investigation into the application of interval and relation theory to the problem of temporal database management*. També disponible la 2a edició a 2014: Time and Relational Theory. San Francisco, US-CA: Morgan Kaufmann, 2002. ISBN: 1-55860-855-9. DOI: 10.1016/B978-155860855-9/50047-7 (citat a les pp. 16, 33, 34, 38, 59, 93).

[34] Jeffrey Dean i Sanjay Ghemawat. "MapReduce: Simplified Data Processing on Large Clusters". A: *Proceedings of the 6th Symposium on Operating Systems Design and Implementation*. OSDI' 04. San Francisco, CA: USENIX Association, 6–8 de des. de 2004, pp. 137 - 150. URL: http://www.usenix.org/legacy/publications/library/proceedings/osdi04/tech/full_papers/dean/dean.pdf (cons. 9-10-2014) (citat a les pp. 36, 197, 199).

[35] Antonios Deligiannakis, Yannis Kotidis i Nick Roussopoulos. "Dissemination of compressed historical information in sensor networks". A: *The VLDB Journal* 16.4 (oct. de 2007), pp. 439 - 461. DOI: 10.1007/s00778-005-0173-5 (citat a les pp. 25, 27, 149).

[36] Alan Demers et al. "The Cougar Project: a work-in-progress report". A: *SIGMOD Record* 32.4 (des. de 2003), pp. 53 - 59. DOI: 10.1145/959060.959070. URL: http://www.cs.cornell.edu/johannes/papers/2003/SigmodRecord2003-Sensor.pdf (cons. 16-4-2012) (citat a les pp. 28, 42).

[37] Luca Deri, Simone Mainardi i Francesco Fusco. "Tsdb: A Compressed Database for Time Series". A: *Proceedings of the 4th International Conference on Traffic Monitoring and Analysis*. TMA '12. Vienna, Austria: Springer-Verlag, des. de 2012, pp. 143 - 156. DOI: 10.1007/978-3-642-28534-9_16. URL: http://luca.ntop.org/tsdb.pdf (cons. 14-11-2013) (citat a la p. 41).

[38] Adam Dou et al. "Supporting Historic Queries in Sensor Networks with Flash Storage". A: *Information Systems* 39 (gen. de 2014), pp. 217 - 232. ISSN: 0306-4379. DOI: 10.1016/j.is.2012.04.002 (citat a la p. 41).

[39] Bradley Dowden. *Time Supplement — The Internet Encyclopedia of Philosophy (IEP)*. 2010. URL: http://www.iep.utm.edu/time-sup/ (cons. 10-12-2013) (citat a la p. 58).

[40] Bradley Dowden. *Time — The Internet Encyclopedia of Philosophy (IEP)*. 2013. URL: http://www.iep.utm.edu/time/ (cons. 6-9-2013) (citat a la p. 58).

[41] Werner Dreyer, Angelika Kotz Dittrich i Duri Schmidt. "An Object-Oriented Data Model for a Time Series Management System". A: *Proceeding of the 7th International Working Conference on Scientific and Statistical Database Management*. Virginia, USA: IEEE Computer Society, set. de 1994, pp. 186 - 195. DOI: 10.1109/SSDM.1994.336948. URL: http://www.ubilab.org/publications/index.html (cons. 30-11-2011) (citat a la p. 39).

[42] Werner Dreyer, Angelika Kotz Dittrich i Duri Schmidt. "Research perspectives for time series management systems". A: *SIGMOD Record* 23.1 (mar. de 1994), pp. 10 - 15. DOI: 10.1145/181550.181553. URL: http://www.ubilab.org/publications/index.html (cons. 30-11-2011) (citat a les pp. 16, 27, 28, 39, 59).

[43] Werner Dreyer, Angelika Kotz Dittrich i Duri Schmidt. *The Implementation of the CALANDA Time Series Management System – A Novel Approach to the Construction of Special-Purpose DBMS*. 1995. URL: http://www.ecofin.ch/aktuelles/presseartikel/zImplementation.pdf (cons. 1-12-2011) (citat a la p. 39).

[44] Werner Dreyer, Angelika Kotz Dittrich i Duri Schmidt. "Using the CALANDA Time Series Management System". A: *SIGMOD Record* 24.2 (mai. de 1995). DOI: 10.1145/568271.223902. URL: http://www.ubilab.org/publications/index.html (cons. 30-11-2011) (citat a les pp. 22, 39).

[45] Keller AG für Druckmesstechnik. *Keller - Logger 5 manual*. Winterthur, CH, oct. de 2012. URL: http://www.catsensors.com/contenidos/productos/Varios-Software/Software/LOGGER/logger5_E_en.pdf (cons. 28-10-2014) (citat a la p. 38).

[46] Stefan Edlich. *NoSQL Databases*. 2014–2009. URL: http://nosql-database.org/ (cons. 13-5-2014) (citat a la p. 36).

[47] Usama Fayyad, Gregory Piatetsky-shapiro i Padhraic Smyth. "From Data Mining to Knowledge Discovery in Databases". A: *AI Magazine* 17.3 (1996), pp. 37 - 54. URL: http://www.aaai.org/ojs/index.php/aimagazine/article/download/1230/1131 (cons. 1-3-2012) (citat a la p. 22).

[48] Zope Foundation. *ZODB - a native object database for Python*. 2011–2009. URL: http://zodb.org/ (cons. 12-5-2014) (citat a la p. 36).

[49] Tak-chung Fu. "A review on time series data mining". A: *Engineering Applications of Artificial Intelligence* 24.1 (feb. de 2011), pp. 164 - 181. DOI: 10.1016/j.engappai.2010.09.007 (citat a les pp. 15, 22).

[50] Wai Fu Fung, David Sun i Johannes Gehrke. "COUGAR: the network is the database". A: *Proceedings of the 2002 ACM SIGMOD international conference on Management of data*. SIGMOD '02. Madison, Wisconsin: ACM, 4 – 6 de jun. de 2002, pp. 621 - 621. DOI: 10.1145/564691.564775. URL: http://www.cs.cornell.edu/johannes/papers/2002/sigmod2002-cougardemo.pdf (cons. 16-4-2012) (citat a la p. 42).

[51] Carlo A. Furia et al. "Modeling time in computing: A taxonomy and a comparative survey". A: *ACM Computing Surveys* 42.2 (feb. de 2010), 6:1 - 6:59. ISSN: 0360-0300. DOI: 10.1145/1667062.1667063 (citat a la p. 92).

[52] Greg Gaughan. *Dee. Makes Python relational*. 2012. URL: http://www.quicksort.co.uk/ (cons. 9-5-2014) (citat a les pp. 35, 252).

[53] Johannes Gehrke i Samuel Madden. "Query processing in sensor networks". A: *Pervasive Computing, IEEE* 3.1 (2004), pp. 46 - 55. DOI: 10.1109/MPRV.2004.1269131 (citat a la p. 25).

[54] Johannes Gehrke et al. *Cougar design and implementation*. Cornell University. 2002. URL: http://www.cs.cornell.edu/boom/2002sp/extproj/www.cs.cornell.edu/database/cougar/cougardesigndoc.pdf (cons. 16-4-2012); *COUGAR: the network is the database*. URL: http://www.cs.cornell.edu/bigreddata/cougar/index.php. Citat a la p. 42.

[55] Pierre Gerard-Marchant i Matt Knox. *scikits.timeseries: Pyton time series analysis*. 2008–2009. URL: http://pytseries.sourceforge.net (cons. 27-10-2014) (citat a la p. 39).

[56] The PostgreSQL Global Development Group. *PostgreSQL: The world's most advanced open source database*. 2012–1996. URL: http://www.postgresql.org/ (cons. 28-5-2012) (citat a la p. 30).

[57] Magnus Lie Hetland. "A survey of recent methods for efficient retrieval of similar time sequences". A: *Data mining in time series databases*. Ed. de Mark Last, Abraham Kandel i Horst Bunke. Series in Machine Perception and Artificial Intelligence 57. Singapore: World Scientific, 2004. Cap. 2, pp. 23 - 41 (citat a les pp. 21, 62, 82, 106).

[58] Icinga. *Icinga - Open Source Monitoring, Nagios fork*. URL: http://www.icinga.org/ (cons. 25-10-2010) (citat a la p. 40).

[59] H. V. Jagadish, Inderpal Singh Mumick i Abraham Silberschatz. "View maintenance issues for the chronicle data model (extended abstract)". A: *Proceedings of the fourteenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems*. PODS '95. San Jose, California, United States: ACM, 1995, pp. 113 - 124. DOI: 10.1145/212433.220201. URL: http:

//www.cs.yale.edu/~avi/home-page/publication-dir/PODS95.pdf (cons. 19-4-2012) (citat a les pp. 28, 164).

[60] H. V. Jagadish et al. "Big Data and Its Technical Challenges". A: *Communications of the ACM* 57.7 (jul. de 2014), pp. 86 - 94. ISSN: 0001-0782. DOI: 10.1145/2611567 (citat a la p. 27).

[61] Neha Jain i Dharma P. Agrawal. "Current Trends in Wireless Sensor Network Design". A: *International Journal of Distributed Sensor Networks* 1.1 (2005), pp. 101 - 122. DOI: 10.1080/15501320590901865 (citat a les pp. 22, 25).

[62] Christian S. Jensen i Richard T. Snodgrass. "Temporal Data Management". A: *IEEE Transactions on Knowledge and Data Engineering* 11.1 (1999), pp. 36 - 44. DOI: 10.1109/69.755613. URL: http://www.cs.arizona.edu/~rts/pubs/TKDEJan99.pdf (cons. 14-12-2012) (citat a les pp. 34, 38).

[63] Christian Søndergaard Jensen i Curtis Dyreson, eds. *The Consensus Glossary of Temporal Database Concepts*. February 1998. Aalborg University. Aalborg Ø, Denmark, 1998. URL: http://people.cs.aau.dk/~csj/Glossary/ (cons. 14-12-2012) (citat a la p. 34).

[64] Eamonn Keogh i Shruti Kasetty. "On the Need for Time Series Data Mining Benchmarks: A Survey and Empirical Demonstration". A: *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '02. Edmonton, Alberta, Canada: ACM, jul. de 2002, pp. 102 - 111. URL: http://www.cs.ucr.edu/~eamonn/sigkdd_bench.pdf (cons. 15-3-2011) (citat a les pp. 23, 252).

[65] Eamonn Keogh i Jessica Lin. *iSAX (Symbolic Aggregate approXimation)*. URL: http://www.cs.ucr.edu/~eamonn/iSAX/iSAX.htm (cons. 15-3-2011) (citat a la p. 40).

[66] Eamonn Keogh i Michael Pazzani. "An enhanced representation of time series which allows fast and accurate classification clustering and relevance feedback". A: *Proceedings of the 4th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '98. New York: ACM, ago. de 1998, pp. 239 - 243. URL: http://www.cs.ucr.edu/~eamonn/kdd98.pdf (cons. 15-3-2011) (citat a la p. 23).

[67] Eamonn Keogh i Padhraic Smyth. "A probabilistic approach to fast pattern matching in time series databases". A: *Proceedings of the 3rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '97. NewPort Beach, California: ACM, ago. de 1997, pp. 24 - 20. URL: http://www.cs.ucr.edu/~eamonn/kdd97.ps (cons. 15-3-2011) (citat a les pp. 16, 23, 25, 27, 233).

[68] Eamonn Keogh et al. "Dimensionality Reduction for Fast Similarity Search in Large Time Series Databases". A: *Knowledge and Information Systems*. KAIS 3.3 (ago. de 2001), pp. 263 - 286. URL: http://www.cs.ucr.edu/~eamonn/kais_2000.pdf (cons. 15-3-2011) (citat a les pp. 23, 40).

[69] Eamonn Keogh et al. "Locally adaptive dimensionality reduction for indexing large time series databases". A: *Proceedings of ACM SIGMOD International Conference on Management of Data*. SIGMOD '01. Santa Barbara, California, USA: ACM, mai. de 2001, pp. 151 - 162. URL: http://www.cs.ucr.edu/~eamonn/sigmod_apca_2001.pdf (cons. 15-3-2011) (citat a la p. 23).

[70] Eamonn Keogh et al. "Segmenting time series: a survey and novel approach". A: *Data mining in time series databases*. Ed. de Mark Last, Abraham Kandel i Horst Bunke. Series in Machine Perception and Artificial Intelligence 57. Singapore: World Scientific, 2004. Cap. 1, pp. 1 - 21 (citat a les pp. 23, 82, 272).

[71] Martin Kersten. *SciQL, Scilens project*. scilens.project.cwi.nl. 2011. URL: http://www.scilens.org/Resources/SciQL (cons. 20-4-2012) (citat a la p. 43).

[72] Martin Kersten et al. "SciQL, a query language for science applications". A: *Proceedings of the EDBT/ICDT 2011 Workshop on Array Databases*. AD '11. Uppsala, Sweden: ACM, 2011, pp. 1 - 12. DOI: 10.1145/1966895.1966896. URL: http://www.cwi.nl/~zhang/papers/arraydb11.pdf (cons. 20-4-2012) (citat a les pp. 28, 43).

[73] Jeong-Joon Kim et al. "Aggregate Queries in Wireless Sensor Networks". A: *International Journal of Distributed Sensor Networks* 2012 (2012). ISSN: 1550-1477. DOI: 10.1155/2012/625798 (citat a la p. 25).

[74] Hermann Kopetz. *Real-Time Systems. Design Principles for Distributed Embedded Applications*. 2a ed. Real-Time Systems Series. New York, US: Springer, 2011. ISBN: 978-1-4419-8236-0. DOI: 10.1007/978-1-4419-8237-7 (citat a les pp. 26, 53, 58, 92, 104, 106).

[75] Bernard Lambeau i University of Louvain. *Alf relational algebra*. 2011–2014. URL: http://www.try-alf.org/ (cons. 16-12-2014) (citat a la p. 252).

[76] Ralf Lämmel. "Google's MapReduce programming model – Revisited". A: *Science of Computer Programming* 70.1 (gen. de 2008), pp. 1 - 30. ISSN: 0167-6423. DOI: 10.1016/j.scico.2007.07.001 (citat a les pp. 80, 197).

[77] Mark Last, Gil Avrahami i Abraham Kandel. "Using data mining techniques for optimizing traffic signal plans at an urban intersection". A: *International Journal of Intelligent Systems* 26.7 (jul. de 2011), pp. 603 - 620. DOI: 10.1002/int.20473 (citat a la p. 22).

[78] Mark Last, Abraham Kandel i Horst Bunke, eds. *Data mining in time series databases*. Series in Machine Perception and Artificial Intelligence 57. Singapore: World Scientific, 2004 (citat a la p. 51).

[79] Mark Last et al. "Knowledge discovery in time series databases". A: *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics* 31.1 (feb. de 2001), pp. 160 - 169. DOI: 10.1109/3477.907576 (citat a les pp. 22, 23, 28, 102, 170, 171).

[80] Camilo Lozoya, Manel Velasco i Pau Martí. "The One-Shot Task Model for Robust Real-Time Embedded Control Systems". A: *IEEE Transactions on Industrial Informatics* 4.3 (jul. de 2008), pp. 164 - 174. DOI: 10.1109/TII.2008.2002702 (citat a la p. 26).

[81] Aleix Llusà Serra. "Disseny i modelització d'un sistema de gestió multiresolució de sèries temporals". Proposta de tesi doctoral. Universitat Politècnica de Catalunya: Programa de doctorat en Automàtica, Robòtica i Visió, jun. de 2012. URL: http://escriny.epsem.upc.edu/attachments/download/13/projecte-tesi.pdf (cons. 30-6-2012) (citat a la p. 51).

[82] Aleix Llusà Serra. "Estudi i modelització dels SGBD Round Robin pel tractament de sèries temporals". Tesi de màster. Universitat Politècnica de Catalunya: Màster Universitari en Automàtica i Robòtica, jun. de 2011. URL: http://lacova.upc.es/~aleix/documents/llusa11-tfm.pdf (cons. 1-1-2014) (citat a la p. 51).

[83] Aleix Llusà Serra. *Implementacions de SGST i de SGSTM. Pytsms, RoundRobinson, RoundRobindoop i Reltsms. Versió 1.0*. 2012–2015. URL: http://escriny.epsem.upc.edu/svn/rrb/src/ (cons. 1-6-2015) (citat a les pp. 18, 183, 236).

[84] Aleix Llusà Serra, Teresa Escobet Canal i Sebastià Vila Marta. "A Model for a Multiresolution Time Series Database System". A: *Proceedings of the 12th International Conference on Artificial Intelligence, Knowledge Engineering and Data Bases*. AIKED '13. Cambridge, UK: WSEAS Press, 20–22 de feb. de 2013, pp. 55 - 60. ISBN: 978-1-61804-162-3. URL: http://www.wseas.us/e-library/conferences/2013/CambridgeUK/AISE/AISE-08.pdf (cons. 29-6-2013) (citat a la p. 18).

[85] Aleix Llusà Serra, Teresa Escobet Canal i Sebastià Vila Marta. *Model and requirements for a multiresolution time series database management system*. Inf. tèc. Universitat Politècnica de Catalunya: Departament de Disseny i Programació de Sistemes Electrònics, 17 de des. de 2012. 20 pp. HDL: 2117/19183 (citat a la p. 18).

[86] Aleix Llusà Serra, Sebastià Vila Marta i Teresa Escobet Canal. "Formalism for a Multiresolution Time Series Database Model". A: *Information Systems* 56 (mar. de 2016), pp. 19 - 35. ISSN: 0306-4379. DOI: 10.1016/j.is.2015.08.006 (citat a la p. 18).

[87] Samuel R. Madden et al. *TinyDB. A declarative database for sensor networks*. 2003–2002. URL: http://telegraph.cs.berkeley.edu/tinydb/index.htm (cons. 17-4-2012) (citat a la p. 42).

[88] Samuel R. Madden et al. "TinyDB: an acquisitional query processing system for sensor networks". A: *ACM Transactions on Database Systems* 30.1 (mar. de 2005), pp. 122 - 173. DOI: 10.1145/1061318.1061322. URL: http://db.cs.berkeley.edu/papers/tods05-tinydb.pdf (cons. 17-4-2012) (citat a les pp. 42, 51).

[89] Sasa Markovic i Arne Vandamme. *JRobin: A Java implementation of RRDtool*. URL: http://www.jrobin.org/ (cons. 22-3-2011) (citat a la p. 40).

[90] Robert Cecil Martin. *Agile Software Development: Principles, Patterns, and Practices*. Prentice Hall, 2002. Cap. Visitor. ISBN: 978-13-597444-5. URL: http://objectmentor.com/resources/articles/visitor.pdf (cons. 4-6-2014) (citat a la p. 187).

[91] Nathan Marz. "A call for sanity in NoSQL". A: *NoSQL Matters Conference*. Barcelona, ES-CA, 29–30 de nov. de 2013. URL: http://2013.nosql-matters.org/bcn/abstracts/#abstract_406338497 (cons. 22-7-2014) (citat a la p. 163).

[92] Nathan Marz i James Warren. *Big data. Principles and best practices of scalable realtime data systems*. Early Access Edition. Manning Publications, 2014. ISBN: 9781617290343. URL: http://www.manning.com/marz/ (cons. 22-7-2014) (citat a la p. 163).

[93] John Mylopoulos et al. "Building knowledge base management systems". A: *The VLDB Journal* 5.4 (1996), pp. 238 - 263. DOI: 10.1007/s007780050027 (citat a la p. 27).

[94] Nagios. *Nagios - The Industry Standard In IT Infrastructure Monitoring*. URL: http://www.nagios.org/ (cons. 25-10-2010) (citat a la p. 40).

[95] Joan Nunes. *Base de dades espacial. Enciclopèdia en línia ICC*. Institut Cartogràfic i Geològic de Catalunya. www.icgc.cat, 2013. URL: http://www.icc.cat/cat/Home-ICC/Mapes-escolars-i-divulgacio/Diccionaris/Base-de-dades-espacial (cons. 8-5-2014) (citat a la p. 34).

[96] Tobias Oetiker. "MRTG The Multi Router Traffic Grapher". A: *Proceedings of the 12th Systems Administration Conference*. LISA '98. Boston, Massachusetts: USENIX Association, des. de 1998, pp. 141 - 148. URL: http://www.usenix.org/publications/library/proceedings/lisa98/oetiker.html (citat a les pp. 40, 98).

[97] Tobias Oetiker. *RRDtool, Round Robin database*. 1998–2013. URL: http://oss.oetiker.ch/rrdtool/ (cons. 21-1-2014) (citat a les pp. 17, 40, 51, 109, 142, 165).

[98] Tobias Oetiker. *The Multi Router Traffic Grapher*. URL: http://oss.oetiker.ch/mrtg/ (cons. 15-10-2010) (citat a la p. 40).

[99] Umit Ogras i Hakan Ferhatosmanoglu. "Online summarization of dynamic time series data". A: *The VLDB Journal* 15.1 (2006), pp. 84 - 98. DOI: 10.1007/s00778-004-0149-x (citat a les pp. 27, 28, 170).

[100] Carl A. Palmer, Nicholos A. Mackos i Michael J. Roemer. "Approach to Monitor and Assess the Quality of Sensor Data in Support of Calibration and Condition Based Maintenance for Turbine Powered Navy Vessels". A: *ASME Conference Proceedings* 1 (2007), pp. 981 - 989. DOI: 10.1115/GT2007-28251 (citat a la p. 22).

[101] Fabian Pascal. *Database Debunkings. Dispelling persistent prevalent database management fallacies*. 2012–2000. URL: http://www.dbdebunk.com/ (cons. 1-4-2012) (citat a les pp. 30, 31, 35).

[102] Andrew Pavlo et al. "A Comparison of Approaches to Large-scale Data Analysis". A: *Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data*. SIGMOD '09. Providence, US-RI: ACM, 2009, pp. 165 - 178. DOI: 10.1145/1559845.1559865. URL: http://database.cs.brown.edu/papers/benchmarks-sigmod09.pdf (cons. 12-5-2014) (citat a la p. 37).

[103] David Plonka, Archit Gupta i Dale Carder. "Application Buffer-Cache Management for Performance: Running the World's Largest MRTG". A: *Proceedings of the 21st Systems Administration Conference*. LISA '07. Dallas, Texas: USENIX Association, nov. de 2007, pp. 63 - 78. URL: http://www.usenix.org/events/lisa07/tech/plonka.html (citat a la p. 40).

[104] John G. Proakis i Dimitris G. Manolakis. *Digital signal processing. Principles, algorithms, and applications*. 3a ed. Upper Saddle River, US-NJ: Prentice-Hall, 1996. ISBN: 978-0-13-394338-9 (citat a la p. 88).

[105] PyData Development Team. *Pandas: python data analysis library*. 2012–2014. URL: http://pandas.pydata.org (cons. 27-10-2014) (citat a la p. 39).

[106] Python Software Foundation. *The Python Standard Library – Python documentation*. Ver. 2.7.7. python.org, 2014–1990. URL: http://docs.python.org/2/library/ (cons. 13-6-2014) (citat a les pp. 185, 187, 193, 207).

[107] Joseba Quevedo et al. "Validation and reconstruction of flow meter data in the Barcelona water distribution network". A: *Control Engineering Practice* 18.6 (jun. de 2010), pp. 640 - 651 (citat a les pp. 22, 25, 26, 53).

[108] Gunter Saake et al. "Downsizing Data Management for Embedded Systems". A: *Egyptian Computer Science Journal (ECS)* 31.1 (gen. de 2009), pp. 1 - 13. URL: http://wwwiti.cs.uni-magdeburg.de/~rosenmue/publications/downsizing09ecs.pdf (cons. 21-5-2012) (citat a la p. 36).

[109] Duri Schmidt et al. "Time Series, A Neglected Issue in Temporal Database Research?" A: *Proceedings of the International Workshop on Temporal Databases: Recent Advances in Temporal Databases*. Workshops in Computing. Zürich, Switzerland: Springer, set. de 1995, pp. 214 - 232. URL: http://www.ecofin.ch/aktuelles/presseartikel/zNeglected_Issue.pdf (cons. 1-12-2011) (citat a les pp. 16, 21, 27, 28, 34, 38, 39).

[110] Arie Segev i Arie Shoshani. "Logical Modeling of Temporal Data". A: *Proceedings of the Association for Computing Machinery Special Interest Group on Management of Data*. SIGMOD '87. San Francisco, US-CA: ACM Press, 27–29 de mai. de 1987, pp. 454 - 466. DOI: 10.1145/38713.38760 (citat a les pp. 16, 28, 39, 88).

[111] Praveen Seshadri. "Enhanced abstract data types in object-relational databases". A: *The VLDB Journal* 7.3 (ago. de 1998), pp. 130 - 140. DOI: 10.1007/s007780050059 (citat a la p. 33).

[112] Praveen Seshadri. "Management of Sequence Data". Tesi doct. University of Wisconsin, 1996. URL: http://www.cs.cornell.edu/home/praveen/papers/thesis.ps.gz (cons. 16-4-2012) (citat a la p. 28).

[113] Praveen Seshadri, Miron Livny i Raghu Ramakrishnan. "SEQ: A model for sequence databases". A: *Proceedings of the Eleventh International Conference on Data Engineering*. ICDE ' 95. Mar. de 1995, pp. 232 - 239. DOI: 10.1109/ICDE.1995.380388. URL: http://www.cs.cornell.edu/home/praveen/papers/seq.de95.ps.Z (cons. 16-4-2012) (citat a la p. 28).

[114] Jin Shieh i Eamonn Keogh. "iSAX: Indexing and Mining Terabyte Sized Time Series". A: *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '08. Las Vegas: ACM, ago. de 2008, pp. 623 - 631. URL: http://www.cs.ucr.edu/~eamonn/iSAX.pdf (cons. 15-3-2011) (citat a les pp. 15, 40).

[115] Francesc Soriguera Martí. "Highway travel time estimation with data fusion". Tesi doct. Barcelona: Universitat Politècnica de Catalunya, oct. de 2010. URL: http://www.fundacioabertis.org/pdf/Editarium_FSoriguera.pdf (cons. 30-4-2012) (citat a la p. 22).

[116] Michael Stonebraker. "Inclusion of New Types in Relational Data Base Systems". A: *Proceedings of the Second International Conference on Data Engineering*. ICDE '86. Los Angeles, California, USA: IEEE Computer Society, feb. de 1986, pp. 262 - 269. URL: http://pages.cs.wisc.edu/~nil/764/ORDB/34_stonebraker86inclusion.pdf (cons. 8-11-2011) (citat a la p. 33).

[117] Michael Stonebraker. "SQL databases v. NoSQL databases". A: *Communications of the ACM* 53.4 (abr. de 2010), pp. 10 - 11. DOI: 10.1145/1721654.1721659 (citat a les pp. 16, 27, 36, 37).

[118] Michael Stonebraker. "Words As Hard As Cannon-balls". A: *NoCOUG Journal* 25.4 (nov. de 2011), pp. 4 - 5. URL: http://www.nocoug.org/Journal/NoCOUG_Journal_201111.pdf (cons. 13-5-2014) (citat a la p. 37).

[119] Michael Stonebraker i Jason Hong. "Saying good-bye to DBMSs, designing effective interfaces". A: *Communications of the ACM* 52.9 (set. de 2009), pp. 12 - 13. DOI: 10.1145/1562164.1562169 (citat a la p. 36).

[120] Michael Stonebraker et al. "Requirements for Science Data Bases and SciDB". A: *Fourth Biennial Conference on Innovative Data Systems Research*. CIDR '09. Asilomar, CA, USA: `www.cidrdb.org`, gen. de 2009. URL: `http://www-db.cs.wisc.edu/cidr/cidr2009/Paper_26.pdf` (cons. 14-7-2014) (citat a les pp. 16, 27, 28, 37, 43).

[121] Michael Stonebraker et al. "The end of an architectural era: (it's time for a complete rewrite)". A: *Proceedings of the 33rd international conference on Very large data bases*. VLDB '07. Vienna, Austria: VLDB Endowment, 2007, pp. 1150 - 1160. URL: `http://www.vldb.org/conf/2007/papers/industrial/p1150-stonebraker.pdf` (cons. 26-3-2012) (citat a les pp. 36, 37).

[122] Abdullah Uz Tansel et al., eds. *Temporal databases: theory, design, and implementation*. Redwood City, CA, USA: Benjamin-Cummings Publishing Co., Inc., 1993. ISBN: 0-8053-2413-5 (citat a la p. 34).

[123] Termcat, Centre de Terminologia, ed. *Diccionaris terminològics del Termcat*. `termcat.cat`, 2014 (citat a la p. 47).

[124] The Apache Software Foundation. *Apache Hadoop*. Ver. 0.20.2. 2010–2014. URL: `http://hadoop.apache.org/` (cons. 12-5-2014) (citat a les pp. 36, 197, 212, 272).

[125] The Graphite Project. *Graphite – scalable realtime graphing*. 2011–2014. URL: `http://graphite.readthedocs.org` (cons. 27-10-2014) (citat a la p. 41).

[126] The Matplotlib Development Team. *Matplotlib documentation*. Ver. 1.3.1. `matplotlib.org`, 2013–2002. URL: `http://matplotlib.org/contents.html` (cons. 13-6-2014) (citat a la p. 188).

[127] The OpenTSDB Authors. *OpenTSDB – A distributed, scalable monitoring system*. 2010–2014. URL: `http://opentsdb.net` (cons. 27-10-2014) (citat a les pp. 43, 252).

[128] Dave Voorhis. *Anonymous and First Class Operators for Tutorial D*. Ver. 1.01. dbappbuilder.sourceforge.net. Ago. de 2012. URL: `http://dbappbuilder.sourceforge.net/docs/AnonymousAndFirstClassOperatorsInTutorialD.pdf` (cons. 26-6-2014) (citat a la p. 218).

[129] Dave Voorhis. *Rel. An Implementation of Date and Darwen's Tutorial D database language*. Ver. 1.0.8. 2012–2004. URL: `http://dbappbuilder.sourceforge.net/` (cons. 1-5-2014) (citat a les pp. 35, 213).

[130] Robert Weigel et al. "TSDS: high-performance merge, subset, and filter software for time series-like data". A: *Earth Science Informatics* 3.1 (2010), pp. 29 - 40. ISSN: 1865-0473. DOI: `10.1007/s12145-010-0059-y` (citat a la p. 42).

[131] Robert Weigel et al. *TSDS (Time Series Database Server)*. 2011 – 2009. URL: `http://tsds.net/` (cons. 2-11-2011) (citat a la p. 43).

[132] Eric W. Weisstein. *Average Function*. MathWorld–A Wolfram Web Resource, created by Eric W. Weisstein. 2013. URL: http://mathworld.wolfram.com/AverageFunction.html (cons. 26-11-2013) (citat a la p. 139).

[133] Wikipedia. *Extended real number line — Wikipedia, The Free Encyclopedia*. URL: http://en.wikipedia.org/wiki/Extended_real_number (cons. 19-9-2011) (citat a la p. 58).

[134] Qiang Yang i Xindong Wu. "10 Challenging Problems in Data Mining Research". A: *International Journal of Information Technology and Decision Making*. IJITDM 5.4 (des. de 2006), pp. 597 - 604. URL: http://www.cs.uvm.edu/~icdm/10Problems/index.shtml (cons. 22-5-2011) (citat a la p. 22).

[135] Yong Yao i Johannes Gehrke. "The Cougar Approach to In-Network Query Processing in Sensor Networks". A: *SIGMOD Record* 31.3 (set. de 2002), pp. 9 - 18. DOI: 10.1145/601858.601861. URL: http://www.cs.cornell.edu/johannes/papers/2002/sigmod-record2002.pdf (cons. 16-4-2012) (citat a les pp. 16, 22, 25).

[136] Byoung-Kee Yi et al. "Online data mining for co-evolving time sequences". A: *Proceedings. 16th International Conference on Data Engineering*. IEEE, 2000, pp. 13 - 22. DOI: 10.1109/ICDE.2000.839383 (citat a la p. 28).

[137] Ming Yu et al. "Prognosis of Hybrid Systems With Multiple Incipient Faults: Augmented Global Analytical Redundancy Relations Approach". A: *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on* 41.3 (mai. de 2011), pp. 540 - 551. ISSN: 1083-4427. DOI: 10.1109/TSMCA.2010.2076396 (citat a la p. 22).

[138] Jian Zhang i Renato Figueiredo. "Adaptive Predictor Integration for System Performance Prediction". A: *Proceedings of the 21st International Parallel and Distributed Processing Symposium*. IPDPS '07. Long Beach, California: IEEE, mar. de 2007. URL: http://www.cecs.uci.edu/~papers/ipdps07/pdfs/IPDPS-1569011105-paper-1.pdf (cons. 3-10-2010) (citat a la p. 40).

[139] Ying Zhang et al. "SciQL: bridging the gap between science and relational DBMS". A: *Proceedings of the 15th Symposium on International Database Engineering & Applications*. IDEAS '11. Lisboa, Portugal: ACM, 2011, pp. 124 - 133. DOI: 10.1145/2076623.2076639. URL: http://www.cwi.nl/~zhang/papers/ideas11.pdf (cons. 20-4-2012) (citat a les pp. 16, 27, 28, 43, 138).

[140] Tarek Ziadé. *Expert Python Programming*. Birmingham, UK: Packt Publishing, set. de 2008. Cap. Useful design patterns. ISBN: 978-1-847194-94-7 (citat a la p. 187).

Abreviacions i nomenclatura

Abreviatures

angl. anglès.

ap. apèndix.

cap. capítol.

fr. francès.

p.ex. per exemple.

§ secció o paràgraf.

s.v. sub voce, 'sota l'entrada' en els diccionaris.

u.t. unitats de temps.

v. vegeu.

Sigles

dbms *Data Base Management System.*

mtsms *Multiresolution Time Series Data Base Management System.*

rdbms *Relational Data Base Management System.*

tsms *Time Series Data Base Management System.*

BDSTM base de dades per a sèries temporals multiresolució (*Multiresolution Time Series Data Base*).

CSV format de fitxer de text que emmagatzema els valors separats per comes (de l'angl. *comma-separated values*).

DD mètode de representació delta de Dirac.

DDL llenguatge de definició de dades (DDL, de l'angl. *data definition language*).

DML llenguatge de manipulació de dades (DML, de l'angl. *data manipulation language*).

FOH mètode de representació first-order hold.

HDFS *Hadoop Distributed File System* [124].

PD mètode de representació parcial discreta.

PLR *Piecewise Linear Representation* [70].

SCADA sistema de supervisió, control i adquisició (de l'angl. *Supervisory Control And Data Acquisition*).

SGBD sistema de gestió de base de dades (*Data Base Management System*).

SGBDR sistema de gestió de base de dades relacional (*Relational Data Base Management System*).

SGST sistema de gestió de base de dades per a sèries temporals (*Time Series Data Base Management System*).

SGSTM sistema de gestió de base de dades per a sèries temporals multiresolució (*Multiresolution Time Series Data Base Management System*).

SIG sistema d'informació geogràfica.

SQL *Structured Query Language*, llenguatge habitual de consulta en els SGBDR.

TAI temps atòmic internacional (del fr. *temps atomique international*).

UML Llenguatge unificat de modelització (de l'angl. *Unified Modeling Language*).

UTC temps universal coordinat (del fr. *temps universel coordonné* i de l'angl. *Coordinated Universal Time*).

VLDB *very large databases*.

XML *Extensible Markup Language*.

ZOH mètode de representació zero-order hold.

ZOHE mètode de representació zero-order hold cap enrere.

Símbols i notació

Índex de notació segons nom, símbol i referència –pàgina o en negreta lloc on es defineix el concepte: definició, secció (§) o exemple (ex.). Organitzada en símbols matemàtics que requereixen aclariment i notació dels components dels SGST i dels SGSTM.

Símbols matemàtics		
A on o tal que	$: \text{o} \mid$	62, 69
Cardinal	$\lvert \ldots \rvert$ (conjunts)	64, 100
Conjunt potència	$\mathcal{P}(\{\})$	79
Domini	$\text{dom}(\{\})$	60, 62, 72–76, 78, 79, 84, 125
Enters	$\mathbb{Z}$	157, 158, 203, 206, 223
F. característica de pertinença	$I_A(t)$	97–99
I lògica	$\wedge$	69, 157
Naturals	$\mathbb{N}$	92, 107, 108, 113, 114, 116, 132, 156, 172, 174, 175, 202, 205
O lògica	$\vee$	72
Producte cartesià	$\times$ (conjunts)	77
Reals	$\mathbb{R}$	57, 58, 60, 61, 63–67, 92, 116, 143, 144, 223, 227, 273, 274
Reals estesos	$\bar{\mathbb{R}}$	57, 58, 61, 63–67, 92, 116, 227, 273
Reals projectius	$\mathbb{R}^*$	60, 61, 63, 143, 144, 274
Separació d'elements	a, b, c (coma i espai)	60, 62
Signe decimal	1,2 (coma sense espai)	89
Valor absolut	$\lvert \ldots \rvert$ (nombres)	100

Notació SGST		**§4**
Temps	$t \in \mathcal{T}$	**§4.1.1**
Domini del temps	$\mathcal{T}$ (p.ex. $\bar{\mathbb{R}}$)	**§4.1.1**, 57, 58, 60, 62–64, 79, 85, 92–95, 97–101, 113, 140, 273

Valor	$v \in \mathcal{V}$	**§4.1.2**
Domini del valor	$\mathcal{V}$ (p.ex. $\mathbb{R}^*$)	**§4.1.2**, 60, 62–64, 79, 81, 94, 152, 274
Mesura	$m = (t, v)$	**4.2**
Valor indefinit	(t, ∞)	**4.5**
Mesura indefinida	$(\pm\infty, \infty)$	**4.6**
Ordre parcial	$m \leq n$	**4.3**
Ordre total	$m \leq^t n$	**4.4**, 61
Temps	$T(m)$	**4.2**
Valor	$V(m)$	**4.2**
Sèrie temporal	$S = \{m_0, \ldots, m_k\}$	**4.7**
Canònica	$(\{t, v\}, \{\ldots\})$	**4.9**
Cardinal	$\|S\|$	**4.8**
Doble	$(\{t_1, v_1, t_2, v_2\}, \{\ldots\})$	**4.11**
Multivaluada	$(\{t, v_1, \ldots, v_n\}, \{\ldots\})$	**4.10**
Operacions de conjunts		**§4.2.1**
Ínfim	$\inf(S)$	**4.17**
Diferència	$S_1 - S_2$	**4.20**
Diferència simètrica	$S_1 \ominus S_2$	**4.24**
Diferència simètrica temporal	$S_1 \ominus^t S_2$	**4.25**, 76
Diferència temporal	$S_1 -^t S_2$	**4.21**, 74
Inclusió	$S_1 \subseteq S_2$	**4.14**
Inclusió temporal	$S_1 \subseteq^t S_2$	**4.15**, 71
Intersecció	$S_1 \cap S_2$	**4.22**
Intersecció temporal	$S_1 \cap^t S_2$	**4.23**, 75
Junció	$S_1 \bowtie S_2$	**4.30**, 78, 81, 119
Màxim	$\max(S)$	**4.16**
Mínim	$\min(S)$	**4.16**
Pertinença	$m \in S$	**4.12**
Pertinença temporal	$m \in^t S$	**4.13**, 69, 71–76
Producte	$S_1 \times S_2$	**4.29**
Projecció	$\Pi_{\{a_0,\ldots,a_n\}}(S)$	**4.27**, 76, 86, 108, 117, 119, 140, 150
Reanomena	$\rho_{a/b}(S)$	**4.28**, 77, 119, 150
Selecció	$\sigma_{a_1 \Theta a_2}(S)$	**4.26**, 76, 81
Suprem	$\sup(S)$	**4.17**
Unió	$S_1 \cup S_2$	**4.18**, 191
Unió temporal	$S_1 \cup^t S_2$	**4.19**, 73, 74, 191
Computacionals		**§4.2.1**

Agregació	agregació(S, m, f)	**4.32**, 80, 81, 138
Binària amb els valors	$S_1 \odot S_2$	**4.35**
Duplica temps	duplica_t(S)	**ex.4.13**, 81, 89
Increments	increments(S)	**ex.4.17**, 82, 89
Mapa	mapa(S, f)	**4.31**, 79–81, 109, 133, 157
Mitjana dels valors	mitjana_v(S)	**ex.4.14**, 81, 109, 140, 142, 174
Màxim dels valors	max_v(S)	**ex.4.14**, 81, 140, 173
Operació binària de valors	$v_1 \odot v_2$	**4.35**, 81, 82
Plec	plec(S, R, f)	**4.33**, 80, 81
Plec amb ordre	oplec(S, R, f, g)	**4.34**, 79, 80, 158
Suma dels valors	suma_v(S)	**ex.4.14**, 174
Valors de les predecessores	vpredecessors(S)	**ex.4.15**, 81, 82
Operacions de seqüències		**§4.2.2**
Concatenació	$S_1 \| \| S_2$	**4.38**, 84, 86, 136, 191
Interval obert	$S(s, t)$	**4.36**
Interval semiobert	$S(s, t]$, $S[s, t)$	**4.36**
Interval tancat	$S[s, t]$	**4.36**
Predecessor	$\text{ant}_S(m)$	**4.37**, 83, 98, 99, 101, 141
Successor	$\text{seg}_S(m)$	**4.37**, 83, 98, 99, 101, 109
Operacions de funció		**§4.2.3**
Concatenació temporal	$S_1 \|\|^r S_2$	**4.41**, 86, 136
Interval temporal	$S[s, t]^r$	**4.39**
Junció temporal	$S_1 \bowtie^r S_2$	**4.42, 4.43**, 86, 134
Selecció temporal	$S[\{t_0, \ldots, t_n\}]^r$	**4.40**
Semijunció temporal	$S_1 \ltimes^r S_2$	**4.43**
Representació	r (nom de mètode)	**§4.3.2**, 95
Delta de Dirac	$r =$ DD	**4.47**, 141, 142, 151, 277
Funció de representació	$S^r(t)$	**4.44**
Graf	graf $S(t)$	**4.45**, 96, 97, 99–101, 103, 279
Parcial discreta	$r =$ PD	**4.46**, 140, 158, 159, 173, 174, 180, 277
Zero-order hold enrere	$r =$ ZOHE	**4.49**, 104, 141, 142, 179, 180, 185, 191, 193, 204, 209, 211, 218, 225, 230, 232, 277

Regularitat		**§4.3.3**
Regular	S és regular	**4.59**
Regulars entre elles	S_1 i S_2 r. entre elles	**4.63**
Disc	$D = (S_D, \kappa)$	**5.3**
Afegeix mesura	afegeixD(D, m)	**5.14**, 124, 132
Màxim cardinal	κ	**5.3**, 114–117, 119–121, 124, 126–129, 132–134, 136, 150, 156–159, 172, 179, 200, 202, 203, 205, 206, 209, 211, 225, 230, 235, 251, 276, 277
Sèrie temporal del disc	S_D	**5.3**

Notació SGSTM		**§5**
Buffer	$B = (S_B, \tau, \delta, f)$	**5.1**
Afegeix mesura	afegeixB(B, m)	**5.11**, 123, 124, 152
Consolida el buffer	consolidaB(B)	**5.12**, 123, 124, 152
Consolidable	B és consolidable	**5.13**, 124
Darrer instant de consolidació	τ	**5.1**
Funció d'agregació d'atributs	f	**5.1**, **§5.3**, 113, 115–117, 119–121, 123, 126–129, 132–136, 150, 152, 156, 172
Pas de consolidació	δ	**5.1**
Sèrie temporal del buffer	S_B	**5.1**
Subsèrie resolució	$R = (B, D)$	**5.5**
Afegeix mesura	afegeixR(R, m)	**5.15**, 124, 125
Consolida	consolidaR(R)	**5.16**, 125
Consolidable	R és consolidable	**5.17**, 125
Sèrie temp. multiresolució	$M = \{R_0, \ldots, R_k\}$	**5.7**
Afegeix mesura	afegeixM(M, m)	**5.18**, 125, 150, 156
Canònica	$(\{S_B, S_D, \tau, \delta, \kappa, f\}, \{\ldots\})$	**5.9**
Consolida	consolidaM(M)	**5.19**, 125, 156
Mapa	mapa(M, f)	**5.20**, 125
Esquema multiresolució	$E = (\{\delta, \tau, f, \kappa\}, \{\ldots\})$	**5.10**, 117, 156–159, 161, 162, 170–172, 179, 200, 202, 203, 205, 206, 209, 225, 230

Afegeix multivalor	afegeixMultivalor(R)	**5.29**, 135
Canvi de mida	canviaK(R, κ)	**5.26**, 132
Canvi de pas de consolidació	canviaδ(R, δ)	**5.27**, 132
Canvi de resolució	canviaResolució(R, k, δ)	**5.28**, 132
Desfasament	desfasamentR(R)	**5.22**, 127–131, 280
Junció de multiresolució	junciómM(M_1, M_2)	**5.33**, 135
Junció de subsèries	junciómR(R_1, R_2)	**5.32**, 134, 135
Lapse de buffer	lapseB(R)	**5.24**, 127–130, 280
Lapse de subsèrie	lapseR(R)	**5.21**, 126–131
Període de buffer	períodeB(R)	**5.23**, 127–131, 280
Subsèrie amb més resolució	maxR(R_1, R_2)	**5.25**, 128
Unió de multiresolució	unióM(M_1, M_2)	**5.31**, 134
Unió de subsèries	unióR(R_1, R_2)	**5.30**, 133, 134
Consultes		**§5.2.3**
Mapa de multiresolució	MapMu($S, \delta, \tau, f, \kappa$) $\equiv$ SèrieDisc	**7.1, 7.2**, 157–159, 162, 166, 200, 204, 280
Plec de multiresolució	PlecMu(S, e) $\equiv$ SèrieTotal	**7.2**, 157, 159, 161, 162, 170–172, 200, 204, 224, 225, 280
Selecció de disc	SèrieDisc(M, δ, f)	**5.34, 5.35**, 157, 162, 192, 193, 211, 230, 277
Sèrie temporal total	SèrieTotal(M)	**5.35**, 136, 137, 157, 158, 161, 162, 170, 192–194, 230, 232, 277, 280, 284
F. d'agregació d'atributs		**§5.3**
Agregació mitjana DD	mitjana$^{\text{DD}}$	**5.37**, 151
Agregació mitjana PD	mitjana$^{\text{PD}}$	**5.36**, 174, 180
Agregació mitjana ZOHE	mitjana$^{\text{ZOHE}}$	**5.38**, 179, 180, 185, 193, 209, 211, 230, 232
Agregació màxim DD	màxim$^{\text{DD}}$	**5.37**, 142, 151
Agregació màxim PD	màxim$^{\text{PD}}$	**5.36**, 158, 159, 173
Agregació màxim ZOHE	màxim$^{\text{ZOHE}}$	**5.38**, 185, 193, 209, 211, 225, 230, 232

Índexs

Índex de figures

Índex de taules

Índex de llistats

Índex de definicions

Índex d'exemples