UNIVERSIDAD DE CANTABRIA

DEPARTAMENTO DE CIENCIAS Y TÉCNICAS
DEL AGUA Y DEL MEDIO AMBIENTE

TESIS DOCTORAL

METODOLOGÍAS DE CALIBRACIÓN DE BASES DE DATOS DE REANÁLISIS DE CLIMA MARÍTIMO

Presentada por: **ANTONIO TOMÁS SAMPEDRO**

Dirigida por: **FERNANDO J. MÉNDEZ INCERA**
IÑIGO J. LOSADA RODRÍGUEZ

Mayo, 2009

CAPÍTULO 3

ESTADO DE CONOCIMIENTO

3.1. Introducción.

Dentro de la metodología general de transferencia de información de oleaje hasta el punto donde se necesita caracterizar el oleaje para un fin concreto (Bases de Datos → Calibración → Clasificación → Propagación → Regímenes), en este capítulo se va a describir el estado de conocimiento actual del segundo punto, la calibración de las bases de datos de oleaje.

Tras describir las bases de datos de oleaje en el capítulo 2, en el presente capitulo 3 se revisan las formas de calibrar las bases de datos que así lo requieran (apartado 3.2). Para ello es necesario explicar las técnicas estadísticas de tratamiento de datos, ya sea las que ajustan unos datos a un modelo determinado, las que ligan o relacionan varias variables entre sí (regresiones) o las que sirven para diagnosticar las propiedades de los distintos conjuntos de datos.

A raíz de la descripción de las bases de datos y de las técnicas estadísticas de tratamiento de datos se comprende la evolución de las metodologías de calibración de bases de datos. Es por ello que en el apartado 3.3 se explica cronológicamente el estado del arte de las metodologías de calibración en función de la aparición de las distintas bases de datos y también del desarrollo de las herramientas estadísticas utilizadas.

Finalmente, en el apartado 3.4 se resume el estado de conocimiento sobre estos temas en una serie de conclusiones y consideraciones, destacando los aspectos más relevantes y característicos y subrayando las carencias o indefiniciones detectadas.

3.2. Técnicas estadísticas de tratamiento de datos.

Como ya se ha descrito, la calibración de bases de datos es un procedimiento de comparación entre varias fuentes de información, de manera que se modifican para tratar de ajustarse, con la mayor exactitud posible, a la realidad. En este apartado se van a desarrollar los métodos o técnicas para modificar o tratar dichas bases de datos.

Cabe puntualizar que las técnicas estadísticas que se van a presentar están basadas fundamentalmente en la estimación, que es una rama de la inferencia estadística. La inferencia estadística está relacionada con los métodos para obtener conclusiones o generalizaciones acerca de una población de datos. Estas conclusiones pueden estar referidas a la forma de la distribución de una variable aleatoria o con los valores de uno o varios de sus parámetros.

El campo de la inferencia estadística se divide en dos, por un lado se tiene el problema de la estimación de los parámetros de una distribución, y por el otro, las pruebas de hipótesis. En el problema de estimación se trata de elegir el valor de un parámetro de la población, mientras que en las pruebas de hipótesis se trata de decidir entre aceptar o rechazar un valor especificado. A su vez, el problema de la estimación se puede dividir en dos áreas: La estimación puntual, y la estimación por intervalos de confianza. En esta tesis se van a utilizar generalmente técnicas de estimación puntual para realizar las calibraciones, aunque en ciertas ocasiones se obtendrán intervalos de confianza de los parámetros estimados. En el esquema de la figura 3.1 se presentan las técnicas de inferencia estadística explicadas.

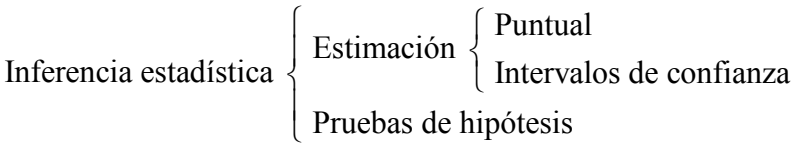

Figura 3.1. Clasificación de las técnicas de inferencia estadística.

Tras esta breve ubicación de la estimación puntual dentro de la inferencia estadística, en este apartado del capítulo 3 se van a describir distintos aspectos y técnicas de tratamiento de datos que se utilizan para calibrar o validar datos oceanográficos. Primeramente se van a definir diversos parámetros estadísticos típicos utilizados a lo largo de la presente tesis. Posteriormente se van a explicar los métodos de búsqueda de estimadores de parámetros con los que mejor se ajustan los datos a un modelo determinado. Después se van a desarrollar los diferentes métodos de regresión, que son técnicas de estimación muy utilizadas para calibrar bases de datos oceanográficas, aunque también pueden utilizarse para comparar, validar o tratar dichos datos. Finalmente, y en relación con todo esto, se van a describir diversas técnicas utilizadas para diagnosticar la bondad de un determinado ajuste, mediante la comparación de distintas poblaciones de datos.

3.2.1. Parámetros estadísticos.

A continuación y antes de explicar las diferentes técnicas estadísticas de tratamiento de datos se van a definir algunos parámetros estadísticos (unidimensionales y bidimensionales) utilizados a lo largo del desarrollo posterior.

3.2.1.1. Parámetros estadísticos unidimensionales.

Sea X una variable aleatoria, de la que se realiza una muestra de n observaciones independientes, $x_1, x_2, ..., x_i, ..., x_n$. La esperanza matemática de la variable aleatoria X se designa por $E[x]$ y es el momento de primer orden respecto del origen.

El momento de orden r respecto de a de la variable X se define como $E[(x-a)^r]$, usualmente cuando a es la media se llaman momentos centrales.

El momento de primer orden respecto del origen ($a=0$) o media poblacional es μ_X, siendo su estimador puntual la media muestral, $\bar{x}$:

$$\mu_X = E[x] = \frac{1}{n}\sum_{i=1}^{n} x_i = \langle x_i \rangle = \bar{x} \tag{3.1}$$

El momento de orden r respecto de la media se define como:

$$\tilde{m}_r = E[(x - E[x])^r] = E[(x-\mu_X)^r] \tag{3.2}$$

Como caso particular de la ecuación 3.2, para $r=2$, el momento de orden 2 o varianza poblacional es ${\sigma_X}^2$, siendo su estimador puntual la cuasivarianza muestral, ${s_X}^2$:

$${\sigma_X}^2 = Var[x] = E\left[\left(x - E[x]\right)^2\right] = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu_X)^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2 = {s_X}^2 \tag{3.3}$$

Definiéndose la desviación típica poblacional como $\sigma_X = \sqrt{{\sigma_X}^2}$ y su estimador puntual que es la cuasidesviación típica muestral, como $s_X = \sqrt{{s_X}^2}$.

Al cociente de la desviación típica entre la media de una variable se denomina coeficiente de variación y se denota por $CV[x] = \sigma_X / \mu_X$.

3.2.1.2. Parámetros estadísticos bidimensionales.

Sea $\{X, Y\}$ una variable aleatoria bidimensional, de la que se realiza una muestra de n observaciones independientes, $\{x_1, y_1\}, \{x_2, y_2\}, ..., \{x_i, y_i\}, ..., \{x_n, y_n\}$. Las varianzas ($S_{XX}$ y S_{YY}) y covarianzas muestrales (S_{XY} y S_{YX}) de definen como:

$$S_{XX} = \sum_{i=1}^{n} (x_i - \bar{x})^2$$
$$S_{YY} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$
$$S_{XY} = S_{YX} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

(3.4)

3.2.2. Ajustes de los datos a un modelo.

Toda calibración paramétrica trata de estimar una serie de parámetros con los que definir la relación de calibración, cuyo vector de parámetros se define como $\Theta = \{\Theta_1, \Theta_2, ..., \Theta_j, ..., \Theta_p\}$, donde p es el número de parámetros. Esta relación de calibración o modelo ($y = g(x;\Theta)$) se determina a partir de una muestra de datos de una o varias variables geofísicas; tratando de encontrar el valor más apropiado para cada uno de los parámetros del modelo. Los valores propuestos para cada parámetro (estimador puntual, $\hat{\Theta}$) se pueden determinar con distintos métodos, siendo los tres más empleados:

- Método de mínimos cuadrados.
- Método de los momentos.
- Método de máxima verosimilitud.

En general, el método de mínimos cuadrados se utiliza para regresiones y los otros dos para ajustar funciones de distribución de probabilidad ($F(x)$). A continuación se van a explicar los tres métodos (Menéndez, 2008), junto con otro basado en el de mínimos cuadrados para ajustar parámetros de funciones de distribución (el método de los papeles probabilísticos, Castillo, 1993).

3.2.2.1. Método de mínimos cuadrados.

El método de mínimos cuadrados (*Ordinary Least-Square*, OLS) ajusta una ecuación (relación paramétrica o modelo de regresión) que liga las distintas variables a una muestra de valores. Así para el modelo de regresión que define las respuestas (y_i) en función de los valores de cada una de las covariables ($x_{i,k}$) e incorpora una serie de parámetros Θ_j, se representa como:

$$y_i = g(x_{i,k};\Theta_j)$$

(3.5)

Se denomina suma de cuadrados residual, SSE (*Sum Squares Error*) a la suma al cuadrado de la diferencia entre la respuesta del modelo de regresión $\hat{y}$ menos la muestra de valores y, de todos los datos de la muestra:

$$SSE = \sum_{i=1}^{n} (y_i - g(x_{i,k}; \Theta_j))^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{3.6}$$

Generalmente lo que se minimiza con el método de mínimos cuadros es SSE, con lo que para este caso la estimación puntual de los parámetros del modelo de regresión ($\hat{\Theta}_j$) se obtendría:

$$\frac{\partial SSE(x;\Theta)}{\partial \Theta_j} = 0; \qquad \text{para } j = 1,..., p \text{ parámetros} \tag{3.7}$$

Cabe señalar que este método de estimación (OLS) normalmente produce estimadores sesgados de los parámetros del modelo.

3.2.2.2. Método de los momentos.

El método de los momentos consiste básicamente en igualar los momentos muestrales con los poblacionales (ver definición en ecuaciones de la 3.1 a la 3.3). Es decir, aceptada una determinada función distribución $F(x;\Theta)$, los momentos de diferente orden calculados a partir de una muestra de datos permiten estimar cuál es el valor del vector de parámetros Θ.

Una extensión del método de los momentos es el método de los momentos de probabilidad ponderados. Al igual que el método de los momentos se asume que los primeros k momentos ponderados de la muestra se aproximan a los momentos ponderados de la población. La principal mejora que aportan los momentos ponderados es que tienen en cuenta la probabilidad asociada a los valores de las colas de las distribuciones. Dichos momentos ponderados vienen definidos por la siguiente expresión:

$$M_{r,s,p} = E[x^r (F(x))^s (1 - F(x))^p] \tag{3.8}$$

Existen tres formas de especial interés $M_{1,s,0}$, $M_{1,0,p}$ y $M_{1,s,p}$; a partir de ciertas combinaciones lineales de los cuales se definen los L-momentos, que contienen información

sobre las características de la función distribución y facilitan las relaciones con los estimadores puntuales.

El método de los momentos es una alternativa razonable para obtener estimadores cuando no se puedan emplear estimadores obtenidos por el método de máxima verosimilitud, que se explica seguidamente.

3.2.2.3. Método de máxima verosimilitud.

El método de máxima verosimilitud (*Maximum Likelihood Estimation*, MLE) consiste en encontrar los estimadores puntuales de los parámetros de una función L que haga máxima la probabilidad de observar los datos de la muestra:

$$L(x;\Theta) = f(x_1,...,x_n;\Theta) = \prod_{i=1}^{n} f(x_i;\Theta) \tag{3.9}$$

La función de verosimilitud $L(x;\Theta)$ es la función de densidad conjunta asociada a todos los valores de partida, con función de densidad $f(x;\Theta)$, siendo Θ el vector de parámetros que caracteriza la función de distribución. Para facilitar los cálculos en la estimación, generalmente se toma el logaritmo neperiano de la ecuación 3.9 y se trabaja con la función logarítmica de verosimilitud, l, tal que $l(x;\Theta) \equiv \log L(x;\Theta)$. La función a maximizar será por tanto:

$$l(x;\Theta) = \sum_{i=1}^{n} \log L(x_i;\Theta) \tag{3.10}$$

Los estimadores máximo-verosímiles son los valores de los parámetros que hacen máxima la probabilidad (verosimilitud), por lo que el valor del estimador puntual máximo-verosímil $\hat{\Theta}_j$ se obtiene resolviendo:

$$\frac{\partial l(x;\Theta)}{\partial \Theta_j} = 0; \qquad \text{para } j = 1,..., p \text{ parámetros} \tag{3.11}$$

Este método proporciona estimadores de mínima varianza (eficientes), pero son generalmente sesgados aunque asintóticamente insesgados.

3.2.2.4. Método de papeles probabilísticos.

El uso de papeles probabilísticos es un método gráfico que permite valorar de una forma sencilla el comportamiento de la muestra respecto a una función distribución, calculando los estimadores puntuales de dicha distribución ($F(x;\Theta)$). Para ello se representa un gráfico bidimensional donde los ejes del gráfico se obtienen mediante transformaciones del tipo $g(x)$ para la variable aleatoria (eje de las abcisas), y $h(y)=h(F(x;\Theta))$ para la probabilidad (eje de las ordenadas), de forma que en dicha métrica la función de distribución es un haz de rectas del tipo $h(y)=\alpha+\beta g(x)$, ver figura 3.2. El ajuste de los datos a la mejor recta mediante el método de los mínimos cuadrados (ver OLS en apartado 3.2.2.1) permite obtener los estimadores puntuales de los parámetros de la recta ($\hat{\Theta}=\{\alpha,\beta\}$) y con ellos los de la función de distribución ajustada.

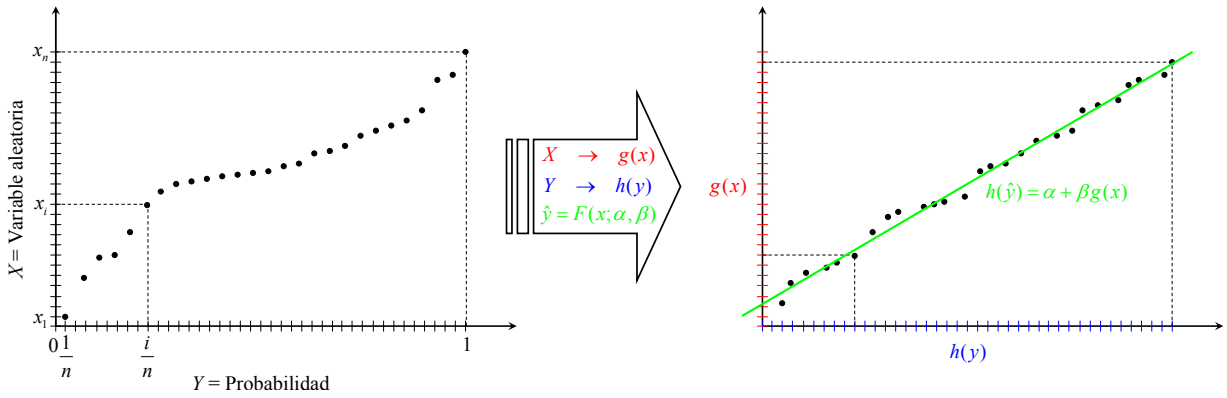

Figura 3.2. Método gráfico de los papeles probabilísticos para la estimación de funciones de distribución.

3.2.3. Regresiones.

Se define como regresión a un modelo de inferencia estadística que relaciona distintas variables e incorpora una serie de parámetros. La estimación de dichos parámetros se realiza mediante el ajuste del modelo a muestras de valores observadas de las variables. Las diferentes hipótesis realizadas a la hora de definir el modelo de regresión y de estimar sus parámetros dan lugar a diferentes tipos de regresiones. A continuación se detallan los más utilizados para la calibración de distintas variables geofísicas de oleaje.

En la figura 3.3 se presentan varias clasificaciones posibles de los tipos de regresiones atendiendo a varios criterios. En primer lugar se pueden clasificar las regresiones en función de las hipótesis en los errores de las distintas variables de la regresión, clásica (con errores en una sola variable), simétrica (alternando que se tenga errores en cada variable y promediando los resultados al final) y EIV (asumiendo errores en todas las variables). También se pueden identificar distintas regresiones en función de si el modelo de regresión es lineal o no. Otra

clasificación se podría realizar atendiendo a si el modelo de regresión está definido con sólo dos variables o con más.

$$\text{Regresiones}\begin{cases}\text{Clásica}\\ \text{Simétrica}\\ \text{EIV}\end{cases} \;\Big\|\; \text{Modelos de regresión}\begin{cases}\text{Lineales}\\ \text{No lineales}\end{cases} \;\Big\|\; \text{Variables}\begin{cases}\text{2 variables}\\ \text{Múltiples}\end{cases}$$

Figura 3.3. Clasificación de los tipos de regresiones en función de varios criterios.

A continuación se van a desarrollar cuatro grandes tipos de regresiones, definiendo para cada una de ellas distintas clases, como son la regresión clásica, la regresión simétrica, la regresión EIV para dos variables y la regresión EIV múltiple muy utilizada en la actualidad, denominada FR. Finalmente, tras la descripción de todos estos tipos de regresiones se van a comparar mediante la aplicación a dos casos de dos variables y a otros dos casos de tres variables.

3.2.3.1. Regresión clásica.

Este método de regresión es ampliamente utilizado y está presente en la mayoría de los libros de estadística (por ejemplo, Luceño, 1989). La regresión clásica es un modelo de inferencia estadística en la que los valores observados de la respuesta del modelo de regresión, $\hat{y}_i$, dependen de los valores de las k covariables, $x_{i,k}$. Siendo Θ_j los parámetros del modelo y e_i errores aleatorios de la respuesta[1], se considera que los errores aleatorios son normalmente distribuidos con media cero y varianza constante. Se asume que las covariables no tienen error y que la varianza de las respuestas es constante (no depende de las covariables), con lo que el modelo de regresión se define como:

$$y_i = g(x_{i,k}, \Theta_j) + e_i = \hat{y}_i + e_i \tag{3.12}$$

La suma de los errores cuadráticos debidos a la desviación de los datos de la curva de regresión ajustada se denomina suma de cuadrados residual, SSE y es la función objetivo a minimizar para estimar los parámetros Θ_j (ver también esta misma ecuación en el epígrafe 3.2.2.1):

[1] Errores aleatorios de la respuesta, e_i : en lugar de utilizar e_{Y_i} para referirse a los errores de la respuesta Y, se utiliza para simplificar la notación e_i, por lo tanto se considerará indistintamente $e_i \equiv e_{Y_i}$.

$$SSE = \sum_{i=1}^{n} e_i^{\,2} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{3.13}$$

En la figura 3.4 se presenta un esquema de la regresión lineal clásica con dos variables (una covariable X y la respuesta Y) en la que el modelo de regresión ($g(x,\Theta)$) es función de la covariable y depende del vector de parámetros Θ. Para estimar el valor de dicho vector de parámetros se minimiza la distancia vertical al cuadrado entre la muestra de las respuestas (y_i) y la respuesta del modelo de regresión ($\hat{y}_i$), es decir se minimiza SSE.

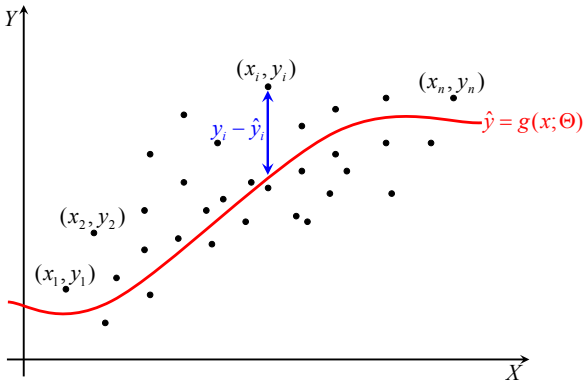

Figura 3.4. Croquis de la regresión clásica.

La suma de errores cuadráticos totales, SST, se compone de SSE (ecuación 3.13) más la suma de errores cuadráticos debidos a la curva de regresión, SSR, de la forma:

$$SST = SSE + SSR; \qquad SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \tag{3.14}$$

A partir de ellos se define el porcentaje de ajuste conseguido con la curva o modelo de regresión, que se denota por R^2 y con él se halla el coeficiente de correlación de la regresión (ρ):

$$R^2 = \frac{SSR}{SST}; \qquad \rho = \sqrt{R^2} \tag{3.15}$$

La variabilidad de los errores o varianza se estima a partir del SSE, de la forma:

$$\sigma_E^{\,2} = Var[e] = \frac{SSE}{n-p} = s_E^{\,2} \tag{3.16}$$

siendo p el número de parámetros a estimar de la regresión.

Finalmente se define el índice de dispersión residual (SI, *Residual Scatter Index*) respecto de la regresión como:

$$SI = \frac{\sqrt{s_E^{\,2}}}{\bar{x}} \tag{3.17}$$

3.2.3.1.1. Regresión lineal (recta).

La regresión lineal clásica (SLR[2], *Simple Linear Regression*) es una regresión clásica en la que el modelo de regresión es lineal respecto de los parámetros (por ejemplo, un polinomio). En general se suele utilizar modelos de regresión que sean rectas (polinomio de grado 1, $p = 2$) cuando se tiene una sola covariable ($k = 1$), por lo que para este caso ($\hat{\Theta} = \{\alpha, \beta\}$):

$$g(x_{i,k}, \Theta_j) = \hat{y}_i = \alpha + \beta x_i \tag{3.18}$$

En la figura 3.5 se presenta un croquis de la regresión lineal clásica para una recta genérica con pendiente β y que corta al eje de las abcisas por α. Para estimar el valor de dichos parámetros se minimiza la distancia vertical al cuadrado entre la muestra de las respuestas (y_i) y la respuesta del modelo de regresión ($\hat{y}_i = \alpha + \beta x_i$).

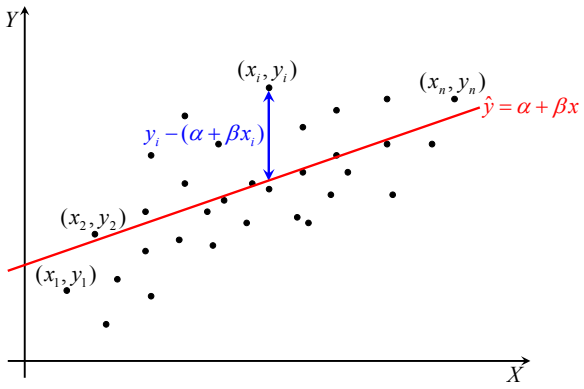

Figura 3.5. Croquis de la regresión lineal clásica para la recta.

Así el valor estimado de los dos parámetros $\hat{\Theta} = \{\alpha, \beta\}$ de la recta se obtiene minimizando SSE por el método OLS (ver apartado 3.2.2.1), definiéndose:

[2] SLR: la regresión lineal clásica se define para muchas variables, pero en general se utiliza SLR para denominar la regresión lineal clásica entre dos variables (una sola covariable).

$$\beta = \frac{S_{XY}}{S_{XX}}$$
$$\alpha = \bar{y} - \beta \bar{x}$$

(3.19)

A partir del valor estimado de los parámetros del modelo de regresión (α y β de la ecuación 3.19) se puede determinar el valor, para este caso de regresión, de los siguientes parámetros de la tabla 3.1:

SSE	SSR	ρ [3]	SI
$S_{YY} - \frac{S_{XY}^{\ 2}}{S_{XX}}$	$\frac{S_{XY}^{\ 2}}{S_{XX}}$	$\frac{S_{XY}}{\sqrt{S_{XX} S_{YY}}}$	$\sqrt{\frac{S_{XX} S_{YY} - S_{XY}^{\ 2}}{S_{XX}(n-2)}} \frac{1}{\bar{x}}$

Tabla 3.1. Parámetros para la regresión lineal $\hat{y}_i = \alpha + \beta x_i$

3.2.3.1.2. Regresión lineal de la recta que pasa por el origen.

El caso más simple de regresión lineal clásica con una sola covariable es el que su modelo de regresión es una recta que pasa por el origen ($p = 1,\ k = 1$ y $\hat{\Theta} = \{\beta\}$):

$$g(x_{i,k}, \Theta_j) = \hat{y}_i = \beta x_i$$

(3.20)

Minimizando SSE por OLS se obtiene el valor estimado del único parámetro de este modelo de regresión ($\hat{\theta} = \{\beta\}$):

$$\beta = \frac{\langle xy \rangle}{\langle x^2 \rangle}$$

(3.21)

En la figura 3.6 se presenta un croquis de la regresión lineal clásica para una recta que corta al eje de las abcisas por el origen, por lo que su parámetro es únicamente la pendiente β. Para estimar el valor se dicho parámetro se minimiza la distancia vertical al cuadrado entre la muestra de las respuestas (y_i) y la respuesta del modelo de regresión ($\hat{y}_i = \beta x_i$).

[3] ρ , coeficiente de correlación de la regresión: para este caso (línea recta de la regresión lineal clásica) también se denomina coeficiente de correlación lineal de Pearson.

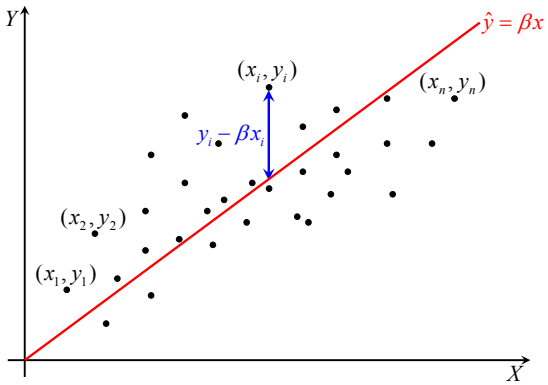

Figura 3.6. Croquis de la regresión clásica para la recta que pasa por el origen.

3.2.3.1.3. Regresión no lineal.

Los casos de regresión clásica presentados anteriormente son lineales. A continuación, se va a presentar un modelo clásico de regresión no lineal típico, basado en una relación potencial del tipo ($p = 2$, $k = 1$ y $\hat{\Theta} = \{\beta, \gamma\}$):

$$g(x_{i,k}, \Theta_j) = \hat{y}_i = \beta x_i^{\gamma} \tag{3.22}$$

Esta regresión se convierte en lineal si se realiza un cambio de variable tomando logaritmos a dicha relación potencial. Una vez realizada la transformación, al ser una regresión lineal clásica, son aplicables las ecuaciones del apartado 3.2.3.1.1 para las variables reducidas.

En la figura 3.7 se presenta un croquis de la regresión clásica no lineal para un modelo de regresión que es una curva potencial del tipo $\hat{y} = \beta x^{\gamma}$, en la que β es la pendiente de la curva en el origen y α controla la curvatura de la función según se aleja del origen. Para estimar el valor de dichos parámetros, al igual que el resto de las regresiones clásicas, se minimiza la distancia vertical al cuadrado entre la muestra de las respuestas y la respuesta del modelo de regresión.

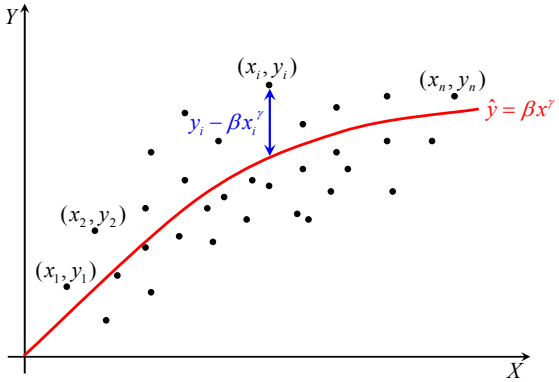

Figura 3.7. Croquis de la regresión clásica no lineal para la relación potencial $\hat{y}_i = \beta x_i^{\gamma}$.

3.2.3.1.4. Regresión lineal múltiple.

La regresión clásica lineal múltiple es una regresión clásica en la que el modelo de regresión es lineal respecto de varias covariables, k, por lo que para este caso el número de parámetros es $p = k + 1$ ($\hat{\Theta} = \{\alpha, \beta_1, \beta_2, ...\beta_k\}$):

$$g(x_{i,k}, \Theta_j) = \hat{y}_i = \alpha + \beta_1 x_{i,1} + \beta_2 x_{i,2} + ... + \beta_j x_{i,j} + ... + \beta_k x_{i,k} \tag{3.23}$$

En forma matricial (que es la notación que se va a seguir para la regresión lineal múltiple) implica:

$$\mathbf{y} = \mathbf{X\Theta} + \boldsymbol{e} \tag{3.24}$$

con:

$$\mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{pmatrix}; \quad \mathbf{X} = \begin{pmatrix} 1 & x_{1,1} & \cdots & x_{1,j} & \cdots & x_{1,k} \\ \vdots & \vdots & & \vdots & & \vdots \\ 1 & x_{i,1} & \cdots & x_{i,j} & \cdots & x_{i,k} \\ \vdots & \vdots & & \vdots & & \vdots \\ 1 & x_{n,1} & \cdots & x_{n,j} & \cdots & x_{n,k} \end{pmatrix}; \quad \mathbf{\Theta} = \begin{pmatrix} \alpha \\ \beta_1 \\ \vdots \\ \beta_j \\ \vdots \\ \beta_k \end{pmatrix}; \quad \boldsymbol{e} = \begin{pmatrix} e_1 \\ \vdots \\ e_i \\ \vdots \\ e_n \end{pmatrix}; \tag{3.25}$$

De forma análoga al caso lineal de una sola covariable, el método de mínimos cuadrados (OLS) proporciona una estimación de los parámetros del modelo, minimizando la siguiente expresión:

$$\boldsymbol{e}^T \boldsymbol{e} = \left(\mathbf{y} - \mathbf{X\Theta}\right)^T \left(\mathbf{y} - \mathbf{X\Theta}\right) \tag{3.26}$$

denotando con el superíndice ^{T} a la matriz transpuesta.

Finalmente el valor estimado de $\hat{\theta}$ se determina a través de la expresión matricial:

$$\hat{\mathbf{\Theta}} = \left(\mathbf{X}^T \mathbf{X}\right)^{-1} \mathbf{X}^T \mathbf{y} \tag{3.27}$$

3.2.3.2. Regresión simétrica.

Otro método de regresión que ha sido utilizado desde Bauer *et al.* (1992) es la regresión simétrica, que debido a sus características ha sido definida para el modelo de regresión que es una recta que pasan por el origen ($Y = \beta X$). Este tipo de regresión realiza la media geométrica de los parámetros estimados con la regresión clásica (asumiendo sólo errores en una variable, la respuesta Y), con los de la una regresión que asume únicamente errores en la otra variable, la covariable X.

Como ya se ha explicado en el apartado 3.2.3.1.2, la regresión clásica de la recta que pasa por el origen ($p = 1$, $k = 1$ y $\hat{\Theta} = \{\beta\}$) se realiza minimizando por OLS el SSE, que es la suma de los errores cuadráticos en la distancia vertical entre y_i e $\hat{y}_i$:

$$SSE = \sum_{i=1}^{n} \left(y_i - \beta_Y x_i \right)^2 \tag{3.28}$$

obteniendo la expresión de la pendiente de dicha recta (igual que la ecuación 3.21):

$$\beta_Y = \frac{\langle xy \rangle}{\langle x^2 \rangle} \tag{3.29}$$

Por el contrario, si se asume que sólo existen errores en la covariable y se minimiza la suma de los errores cuadráticos en la distancia horizontal entre x_i e $\hat{x}_i$:

$$\sum_{i=1}^{n} \left(x_i - \frac{y_i}{\beta_X} \right)^2 \tag{3.30}$$

por el método OLS se obtiene:

$$\beta_X = \frac{\langle y^2 \rangle}{\langle xy \rangle} \tag{3.31}$$

En la figura 3.8 se presenta la interpretación gráfica de ambas regresiones. Cuando se minimiza la distancia vertical (regresión clásica) se obtiene la expresión de β_Y de la ecuación 3.29. En cambio si se minimiza la distancia horizontal se determina la expresión de β_X de la ecuación 3.31.

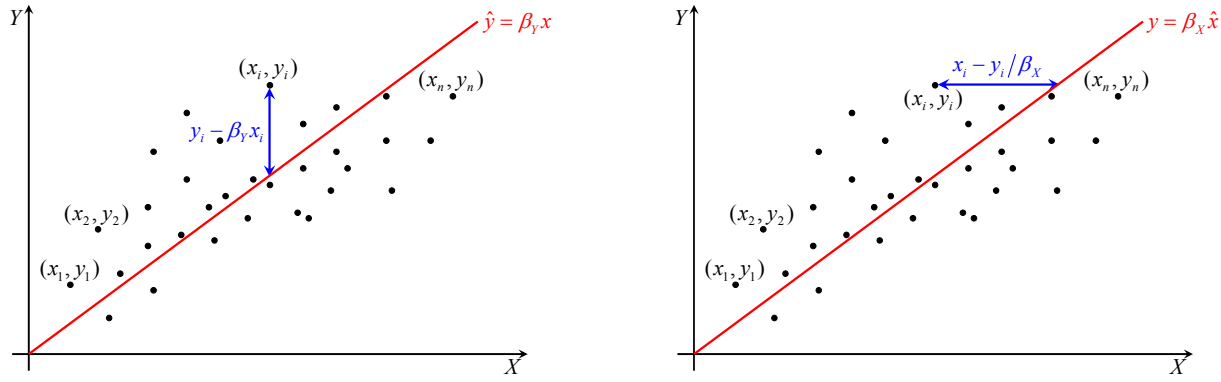

Figura 3.8. Regresión para la recta que pasa por el origen, minimizando la distancia vertical (izquierda) y minimizando la distancia horizontal (derecha).

Un inconveniente de las definiciones de β_Y y β_X de las ecuaciones 3.29 y 3.31 respectivamente es que están sesgadas. Debido a la desigualdad de Cauchy-Schwartz:

$$\frac{\langle xy \rangle^2}{\langle x^2 \rangle \langle y^2 \rangle} \leq 1 \tag{3.32}$$

esto implica que $\beta_Y \leq \beta_X$.

Por ello se define el coeficiente de regresión simétrico, como la media geométrica de ambos parámetros, siendo independiente del intercambio de X e Y:

$$\beta = \sqrt{\beta_X \beta_Y} = \sqrt{\frac{\langle y^2 \rangle}{\langle x^2 \rangle}} \tag{3.33}$$

Dicho parámetro está definido siempre como positivo, por lo que habrá que tener en cuenta de manera independiente el signo que debe tener el modelo de regresión.

3.2.3.3. Regresión EIV.

El modelo de regresión EIV, *Errors-In-Variables* considera que todas las variables han sido medidas con error; siendo más general que la regresión clásica que asume únicamente errores en la respuesta del modelo de regresión. Por tanto, la regresión EIV tiene una mayor aplicabilidad en diferentes modelos de regresión que la regresión simétrica. En este apartado se van a desarrollar distintos tipos de regresiones EIV de dos variables, explicando las regresiones EIV múltiples en el apartado posterior.

Así, de dos variables X e Y (matemáticas o deterministas) que no pueden observarse directamente se pueden obtener unas muestras ($\hat{X}$ e $\hat{Y}$ respectivamente) que sí tienen errores. Por lo tanto e_X y e_Y son los errores de $\hat{X}$ e $\hat{Y}$ con respecto a las variables deterministas X e Y de la forma:

$$\hat{X} = X + e_X$$
$$\hat{Y} = Y + e_Y \tag{3.34}$$

Se asume que las mediciones de los errores e_X y e_Y se distribuyen normalmente con media cero y varianza constante, expresándose las varianzas de los errores como:

$$Var[e_X] = {\sigma_{EX}}^2$$
$$Var[e_Y] = {\sigma_{EY}}^2 \tag{3.35}$$

Una vez definidos los modelos de regresión paramétricos que ligan X e Y, se pueden utilizar varios procedimientos para estimar los parámetros en el modelo de regresión EIV. Cuando se utiliza la estimación por mínimos cuadrados, OLS (ver OLS en apartado 3.2.2.1) se obtienen estimadores sesgados de los parámetros, pero cuando ${\sigma_{EX}}^2$ es pequeño frente a ${\sigma_{EY}}^2$ el sesgo es pequeño, pudiendo llegar a anularse si X está predeterminada (hipótesis de la regresión clásica). Si se realiza la estimación por el método de máxima verosimilitud, MLE (ver MLE en apartado 3.2.2.3), es necesaria información adicional para estimar dichos parámetros. Por ello, Wong (1989) define el parámetro λ como la fracción de las varianzas de los errores en las mediciones:

$$\lambda = \frac{{\sigma_{EY}}^2}{{\sigma_{EX}}^2} \tag{3.36}$$

Si $\lambda \to \infty$ se obtienen las expresiones estimadas por la regresión clásica con OLS. Pero normalmente λ es desconocido pues ni ${\sigma_{EX}}^2$ ni ${\sigma_{EY}}^2$ son conocidos, por lo que surgen dos versiones en función de las distintas hipótesis para definir λ:

- ODR, *Orthogonal Distance Regression,* asume $\lambda = 1$.
- GMFR, *Geometric Mean Functional Relationship*, supone $\lambda = \frac{S_{YY}}{S_{XX}}$.

Cabe señalar que Marsden (1999) comparando ajustes utilizando la regresión clásica con los de ambos tipos de la regresión EIV, y para ajustes con coeficientes de correlación altos ($\rho > 0.9$), obtuvo los mismos resultados; no siendo así para correlaciones bajas. A continuación se definen las regresiones ODR y GMFR para distintos modelos de regresión.

3.2.3.3.1. Regresión ODR.

Esta regresión, tomada de Boggs y Rogers (1990), asume $\lambda = 1$ (por ejemplo ${\sigma_{EX}}^2 = {\sigma_{EY}}^2$) y se denomina regresión de distancia ortogonal porque minimiza la distancia ortogonal entre las observaciones y la curva de ajuste, a diferencia de la regresión clásica que minimiza la distancia vertical. Dicha línea es también la primera componente principal de los datos, por eso es igualmente llamada *Principal Component Regression*[4].

En la figura 3.9 se presenta un esquema de la regresión ODR, que minimiza la distancia ortogonal de la muestra de puntos al modelo de regresión por el método MLE. Se muestra tanto para el modelo de regresión recta ($Y = \alpha + \beta X$), como para la recta que pasa por el origen ($Y = \beta X$). Ambos se desarrollan a continuación, pero la aplicación de la regresión ODR a otros casos más complicados o no lineales es relativamente sencilla.

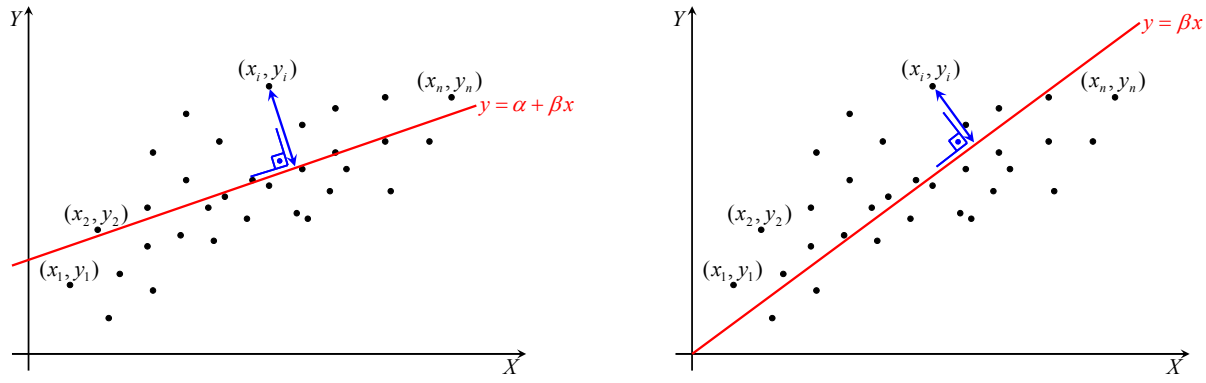

Figura 3.9. Croquis de la regresión ODR para la recta (izquierda) y la recta que pasa por el origen (derecha).

3.2.3.3.1.1. Recta.

De las regresiones ODR, en este epígrafe se va a explicar el modelo de regresión lineal que es una recta (con dos parámetros a estimar, $\hat{\Theta} = \{\alpha, \beta\}$), tomado de Soukissian y Kechris (2007). Esta regresión es análoga a la regresión lineal clásica, SLR (ver SLR en apartado 3.2.3.1.1) pero con errores tanto en la variable aleatoria X como en la Y, quedando definido por:

[4] *Principal Component Analysis*, PCA: se puede ver el desarrollo del análisis de componentes principales a variables geofísicas en el anejo II (EOF).

$$Y = \alpha + \beta X; \qquad \hat{Y} - e_Y = \alpha + \beta(\hat{X} - e_X) \tag{3.37}$$

Minimizando por el método MLE el modelo de regresión propuesto (recta) se llega a la expresión que depende del parámetro λ (ver ecuación 3.36):

$$\beta = \frac{S_{YY} - \lambda S_{XX} + \sqrt{(S_{YY} - \lambda S_{XX})^2 + 4\lambda S_{XY}{}^2}}{2S_{XY}}$$
$$\alpha = \bar{y} - \beta \bar{x} \tag{3.38}$$

Si se introduce en esta expresión la hipótesis del ODR ($\lambda = 1$) se obtiene:

$$\beta = \frac{S_{YY} - S_{XX} + \sqrt{(S_{YY} - S_{XX})^2 + 4S_{XY}{}^2}}{2S_{XY}}$$
$$\alpha = \bar{y} - \beta \bar{x} \tag{3.39}$$

3.2.3.3.1.2. Recta que pasa por el origen.

Entre las regresiones ODR también se ha aplicado a modelos de rectas que pasan por el origen ($Y = \beta X$), como por ejemplo Bauer *et al.* (1992). Este modelo de regresión también se ha utilizado para las regresiones clásica y simétrica, pero ahora se asumen errores simultáneamente tanto en la variable aleatoria X como en la Y, quedando definido por:

$$Y = \beta X; \qquad \hat{Y} - e_Y = \beta(\hat{X} - e_X) \tag{3.40}$$

Para estimar el único parámetro de la regresión ($\hat{\Theta} = \{\beta\}$) se minimiza por el método MLE el modelo de regresión propuesto, con la hipótesis ODR de $\sigma_{EX}{}^2 = \sigma_{EY}{}^2$, llegando a la expresión:

$$\beta = \tan\left[\frac{1}{2}\tan^{-1}\left(\frac{\langle 2xy \rangle}{\langle x^2 \rangle - \langle y^2 \rangle}\right)\right] \tag{3.41}$$

Esta expresión es muy sensible cuando el coeficiente de correlación es próximo a uno. Esto es debido a que el denominador varía fácilmente de signo cuando X e Y son muy próximos. Para solucionar estos problemas Hwang *et al.* (1998) definió β como:

$$\beta = \begin{cases} \tan\left[\frac{1}{2}\tan^{-1}\left(\frac{\langle 2xy\rangle}{\langle x^2\rangle - \langle y^2\rangle}\right)\right]; & \langle x^2\rangle \geq \langle y^2\rangle \\ \tan\left[\frac{1}{2}\left\{\tan^{-1}\left(\frac{\langle 2xy\rangle}{\langle x^2\rangle - \langle y^2\rangle}\right) + \pi\right\}\right]; & \langle x^2\rangle < \langle y^2\rangle \end{cases} \tag{3.42}$$

3.2.3.3.2. Regresión GMFR.

De los dos tipos de regresiones EIV (ODR y GMFR) en este apartado se explica la regresión GMFR aplicada al modelo de regresión lineal que es una recta (con dos parámetros a estimar, $\hat{\Theta} = \{\alpha, \beta\}$), tomado de Soukissian y Kechris (2007); este modelo de regresión es idéntico al planteado anteriormente para la regresión recta ODR (ver ecuación 3.37):

$$Y = \alpha + \beta X; \qquad \hat{Y} - e_Y = \alpha + \beta(\hat{X} - e_X) \tag{3.43}$$

Y de igual manera se obtienen en función de λ, por el método MLE (ver ecuación 3.38), la estimación de los parámetros de la recta:

$$\beta = \frac{S_{YY} - \lambda S_{XX} + \sqrt{(S_{YY} - \lambda S_{XX})^2 + 4\lambda S_{XY}{}^2}}{2S_{XY}}$$
$$\alpha = \overline{y} - \beta\overline{x} \tag{3.44}$$

Utilizando ahora la hipótesis del GMFR, $\lambda = \frac{S_{YY}}{S_{XX}}$, la expresión de ambos parámetros queda:

$$\beta = \sqrt{\frac{S_{YY}}{S_{XX}}}$$
$$\alpha = \overline{y} - \beta\overline{x} \tag{3.45}$$

La denominada relación funcional de la media geométrica (GMFR) debe su nombre a que el estimador de β es la media geométrica de las pendientes de las regresiones de Y sobre X y de X sobre Y. En general, si no se tiene información para elegir otra, se recomienda utilizar la regresión GMFR (Draper y Smith, 1998).

3.2.3.4. Regresión FR (EIV múltiple).

El modelo de regresión denominado relación funcional o estructural (FR, *Functional Relationship*) es un tipo de regresión EIV múltiple en la que se considera que una de las variables es determinista, es decir se considera que es la realidad (en adelante a esta variable se la denota por T, *Truth*). En algunas ocasiones a este tipo de regresión también se la llama regresión neutra (*Neutral Regression*) por la forma de estimar los parámetros (Marsden, 1999). Una diferencia fundamental con respecto a los modelos EIV es que no hay que realizar hipótesis sobre el error en las mediciones de cada variable, el método FR es capaz de estimar las varianzas de los errores de cada variable (no hay que hacer hipótesis sobre el valor de λ como en los casos EIV presentados).

Para aplicaciones en bases de datos de oleaje, en general se suelen utilizar modelos múltiples de tres variables (X, Y y Z) que estiman la realidad con errores aleatorios distribuidos normalmente de media cero y varianza constante (e_X, e_Y y e_Z). Normalmente, en el caso que nos ocupa, los datos de las tres variables son de modelado numérico, boyas y satélites. Las distintas relaciones de las tres variables o parametrizaciones del modelo FR, dan lugar a distintos modelos de regresión. A continuación, se presentan algunos de los más usuales en calibraciones de datos de oleaje.

3.2.3.4.1. Modelo con rectas en dos de las tres variables.

Este modelo de regresión ha sido tomado de Caires y Sterl (2003) y relaciona la variable real, T, con tres variables observadas (X, Y y Z con errores e_X, e_Y y e_Z respectivamente) por medio de dos rectas en dos de ellas e identificando directamente T con la tercera:

$$\begin{aligned} X &= T + e_X \\ Y &= \alpha_1 + \beta_1 T + e_Y \\ Z &= \alpha_2 + \beta_2 T + e_Z \end{aligned} \tag{3.46}$$

En la relación funcional anterior (ecuación 3.46) los parámetros de la regresión, $\hat{\Theta} = \{\alpha_1, \alpha_2, \beta_1, \beta_2\}$, se estiman mediante las expresiones siguientes:

$$\begin{aligned}\beta_1 &= S_{YZ}/S_{XZ}\\ \beta_2 &= S_{YZ}/S_{XY}\\ \alpha_1 &= \overline{y} - \beta_1 \overline{x}\\ \alpha_2 &= \overline{z} - \beta_2 \overline{x}\end{aligned} \tag{3.47}$$

Y las varianzas de los errores se estiman mediante:

$$\begin{aligned}\sigma_{EX}{}^2 &= \tfrac{1}{n} S_{XX} - \tfrac{1}{n} S_{XY} S_{XZ}/S_{YZ}\\ \sigma_{EY}{}^2 &= \tfrac{1}{n} S_{YY} - \tfrac{1}{n} S_{XY} S_{YZ}/S_{XZ}\\ \sigma_{EZ}{}^2 &= \tfrac{1}{n} S_{ZZ} - \tfrac{1}{n} S_{XZ} S_{YZ}/S_{XY}\end{aligned} \tag{3.48}$$

Una ventaja de este tipo de regresiones es su naturaleza simétrica, que significa que el resultado de aplicar el modelo a los datos es independiente del orden de la identificación de las variables a X, Y o Z. Esto implica que se puede determinar los coeficientes de la relación cruzada entre dos de las variables, pudiendo obtener relaciones de calibración como por ejemplo $Y = \alpha_3 + \beta_3 Z$ a partir de las ecuaciones de 3.47:

$$\begin{aligned}\beta_3 &= \beta_1/\beta_2 = S_{XY}/S_{XZ}\\ \alpha_3 &= \alpha_1 - \alpha_2\ \beta_1/\beta_2 = \overline{y} - \beta_3 \overline{z}\end{aligned} \tag{3.49}$$

3.2.3.4.2. Modelo con rectas que pasan por el origen en dos de las tres variables.

Este modelo de regresión es más sencillo (su relación funcional tiene menos parámetros) que el expuesto previamente, pues se relaciona la variable real, T, con las tres variables observadas (X, Y y Z con errores e_X, e_Y y e_Z respectivamente) por medio de dos rectas que pasan por el origen en dos de ellas y se identifica directamente T con la tercera, siendo:

$$\begin{aligned}X &= T + e_X\\ Y &= \beta_1 T + e_Y\\ Z &= \beta_2 T + e_Z\end{aligned} \tag{3.50}$$

Se utilizan únicamente dos parámetros, $\hat{\Theta} = \{\beta_1, \beta_2\}$, que son las pendientes de las rectas de regresión pasando por el origen, y se estiman por:

$$\begin{aligned} \beta_1 &= \langle yz \rangle / \langle xz \rangle \\ \beta_2 &= \langle yz \rangle / \langle xy \rangle \end{aligned} \tag{3.51}$$

Las varianzas de los errores se estiman mediante las expresiones:

$$\begin{aligned} \sigma_{EX}{}^2 &= \langle x^2 \rangle - \langle xy \rangle \langle xz \rangle / \langle yz \rangle \\ \sigma_{EY}{}^2 &= \langle y^2 \rangle - \langle xy \rangle \langle yz \rangle / \langle xz \rangle \\ \sigma_{EZ}{}^2 &= \langle z^2 \rangle - \langle xz \rangle \langle yz \rangle / \langle xy \rangle \end{aligned} \tag{3.52}$$

Al igual que en el caso anterior (epígrafe 3.2.3.4.1), por la naturaleza simétrica de estas regresiones, la pendiente de la recta $Y = \beta_3 Z$ es determinada por:

$$\beta_3 = \beta_1 / \beta_2 = \langle xy \rangle / \langle xz \rangle \tag{3.53}$$

3.2.3.4.2. Modelo con rectas que pasan por el origen en las tres variables.

El modelo de regresión tomado de Janssen *et al.* (2003) relaciona la variable real, T, con las tres variables observadas (X, Y y Z con errores e_X, e_Y y e_Z respectivamente) por medio de tres rectas que pasan por el origen, definiendo la relación funcional de la siguiente forma:

$$\begin{aligned} X &= \beta_1 T + e_X \\ Y &= \beta_2 T + e_Y \\ Z &= \beta_3 T + e_Z \end{aligned} \tag{3.54}$$

Las varianzas de los errores se determinan mediante las expresiones:

$$\sigma_{EX}{}^2 = \langle (x - \beta_1\, y/\beta_2)(x - \beta_1\, z/\beta_3) \rangle$$
$$\sigma_{EY}{}^2 = \langle (y - \beta_2\, x/\beta_1)(y - \beta_2\, z/\beta_3) \rangle$$
$$\sigma_{EZ}{}^2 = \langle (z - \beta_3\, x/\beta_1)(z - \beta_3\, y/\beta_2) \rangle \tag{3.55}$$

Las expresiones de las varianzas de los errores dependen de los parámetros del modelo ($\hat{\Theta} = \{\beta_1, \beta_2, \beta_3\}$), por lo que no puede obtenerse la estimación de todos los parámetros. Sin pérdida de generalidad, se puede tomar X como referencia y calcular las constantes de calibración de Y y Z, siendo la de Y:

$$\beta_2 = \left(-B + \sqrt{B^2 - 4AC}\right) \Big/ (2A) \tag{3.56}$$

con $A = \langle xy \rangle / \lambda$, $\lambda = \sigma_{EY}{}^2 / \sigma_{EX}{}^2$, $B = \langle x^2 \rangle - \langle y^2 \rangle / \lambda$ y $C = -\langle xy \rangle$. Sustituyendo Y por Z se obtiene β_3. Como las constantes de calibración dependen de las varianzas del error se pueden calcular de manera iterativa, comenzando por tomar las pendientes igual a 1 y en menos de 10 iteraciones se converge a la solución final.

3.2.3.5. Aplicación de regresiones con dos variables.

Tras haber presentado los tipos de regresiones más usualmente utilizados para calibrar bases de datos de oleaje, en este apartado se va a aplicar dichos métodos de regresión a dos casos. Ambos casos tienen dos variables, por lo que aquí no se van a presentar ejemplos de las aplicaciones de las regresiones múltiples (en el apartado 3.2.3.6 se presentarán dos ejemplos de aplicación de regresión múltiple FR con tres variables). En la figura 3.10 se recapitulan todas las regresiones mostradas, subrayando (en el color que aparecerán posteriormente) los tipos de regresiones bidimensionales.

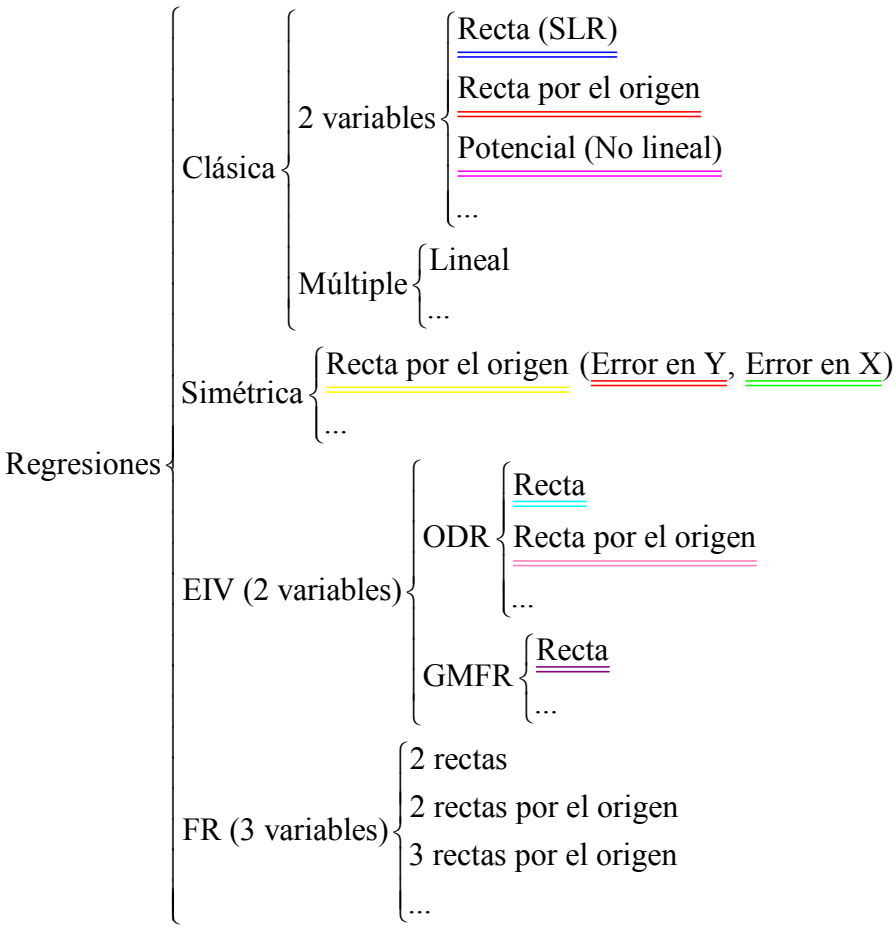

Figura 3.10. Clasificación de las regresiones explicadas, subrayando los tipos de dos variables.

Se van a comparar las distintas regresiones de dos variables explicadas mediante la aplicación a dos ejemplos. Los datos utilizados son la H_s proveniente de dos bases de datos para dos ubicaciones distintas. La primera base de datos es el reanálisis de SIMAR-44 perteneciente al OPPE, que será la variable Y y la otra base de datos son las boyas de la red Exterior del OPPE, que será la variable X. Las dos ubicaciones donde se realizan las comparaciones son la posición de la boya de Villano-Sisargas (Atlántico) y la de Cabo de Gata (Mediterráneo). En la tabla 2.1 y en la figura 2.15 del capítulo 2 se pueden ver dichas posiciones. En las figuras que se van a mostrar a continuación, se representan los datos de SIMAR en el eje de abcisas frente a los de la boya en el eje de ordenadas, para cada una de las posiciones de las boyas.

Al representar los datos de cada boya frente a los coincidentes en tiempo y posición de SIMAR-44, si fuesen exactamente iguales deberían situarse sobre la recta bisectriz. Como puede verse en la figura 3.11, para Villano-Sisargas los puntos se sitúan bastante próximos a la bisectriz, en cambio para Cabo de Gata no ocurre lo mismo, presentando mucha más dispersión (estos dos puntos representan 2 casos extremos que nos vamos a encontrar a lo largo de la tesis). Podría decirse que ambas bases de datos (reanálisis y boyas) son más

parecidas en la zona de Villano-Sisargas que en la de Cabo de Gata, a pesar de que los datos de Villano-Sisargas ($0\ \text{m} \le H_s \le 12\ \text{m}$) son más energéticos que los de Cabo de Gata ($0\ \text{m} \le H_s \le 5\ \text{m}$).

En la figura 3.11 se representa también la recta de ajuste de la regresión lineal clásica (SLR); observándose como para Villano-Sisargas la recta es casi coincidente con la recta bisectriz, α es casi 0 y β es casi 1; en cambio para Cabo de Gata ambas rectas son muy diferentes. Posteriormente se va a representar diversas figuras con distintas regresiones, pero en todas ellas se va seguir mostrando las rectas SLR, pues sirve de referencia para comparar ya que dicha regresión ha sido tradicionalmente la más utilizada.

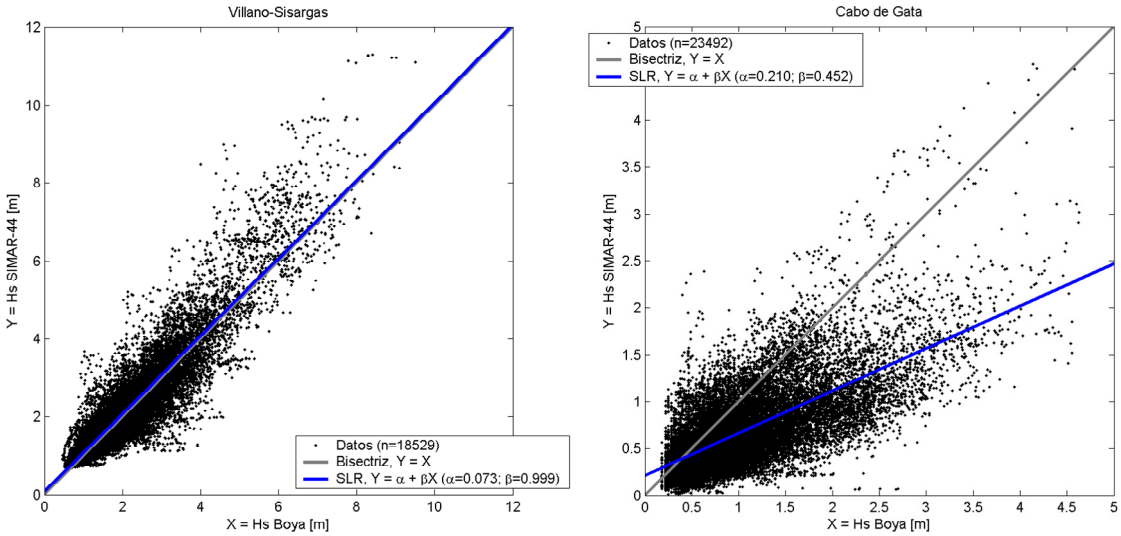

Figura 3.11. Ajuste de recta SLR a datos de H_s de dos boyas y sus correspondientes de SIMAR-44.

La recta que pasa por el origen ($Y = \beta X$) es un modelo de regresión ampliamente utilizado para calibrar y permite contrastar fácilmente sus ajustes con distintos tipos de regresiones, pues sólo depende de un parámetro, β. En la figura 3.12 se representan distintos ajustes de dicho parámetro para las regresiones anteriormente descritas aplicadas a los datos de H_s de Villano-Sisargas y Cabo de Gata, es decir: para la clásica o minimizando la distancia vertical (β_Y); minimizando la distancia horizontal (β_X), la simétrica (β_{SIM}) y la que minimiza la distancia ortogonal (β_{ODR}). Los distintos ajustes se aproximan más en Villano-Sisargas que en Cabo de Gata debido a que tiene una menor dispersión de los datos; pero puede observarse como las pendientes β_{SIM} y β_{ODR} están siempre acotadas entre β_Y y β_X, siendo muy próximas entre sí (para Villano-Sisargas prácticamente coincidentes), pues ambas consideran errores iguales en X y en Y. Para las pendientes más extremas, β_Y es menor que β_X, pero esto está motivado por la distribución de los datos de ambos ejemplos.

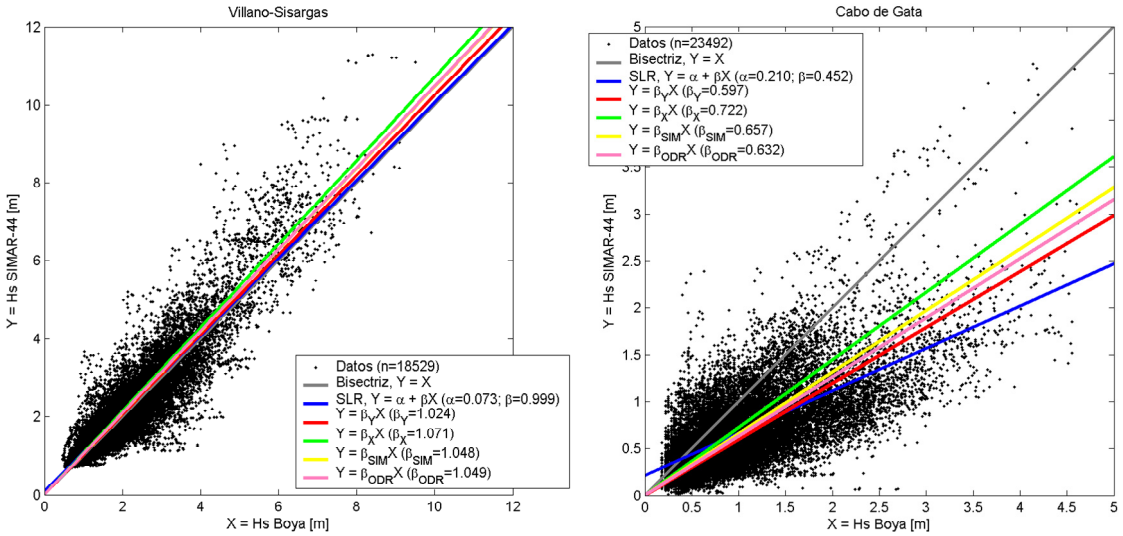

Figura 3.12. Ajuste de diferentes rectas que pasan por el orinen ($Y = \beta X$) y recta SLR con datos de H_s de dos boyas y sus correspondientes de SIMAR-44.

En la figura 3.13 se incorpora el ajuste de la regresión clásica al modelo no lineal de regresión $Y = \beta X^{\gamma}$, obteniendo una curva similar a la recta SLR para los valores medios, pero partiendo del origen. El ajuste de estas curvas potenciales no recogen los distintos comportamientos de los datos medios del oleaje y los más extremos, pues hay mucha más población de datos medios y los ajustes se aproximan más a ellos.

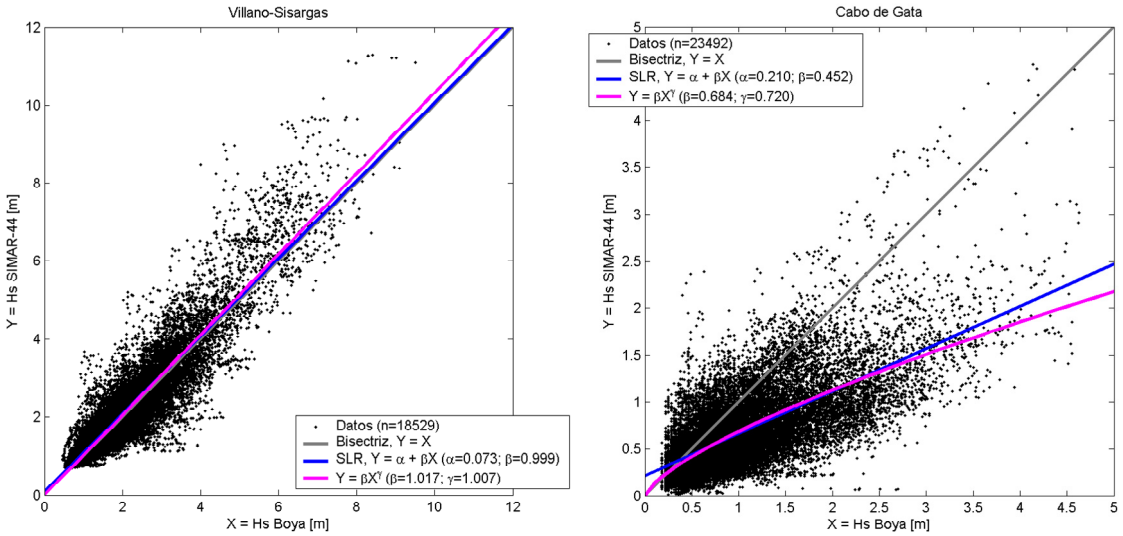

Figura 3.13. Ajuste de relación no lineal ($Y = \beta X^{\gamma}$) y recta SLR a datos de H_s de dos boyas y sus correspondientes de SIMAR-44.

Finalmente se representan en la figura 3.14 todas las regresiones EIV (ODR y GMFR) detalladas anteriormente. Puede comprobarse que como se comentó anteriormente las rectas $Y = \alpha + \beta X$, ODR y GMFR, prácticamente coinciden para correlaciones altas (caso Villano-

Sisargas), siendo más distintas a medida que disminuye la correlación. También se aprecia como las regresiones EIV distan más de la SLR para dispersiones altas (caso Cabo de Gata).

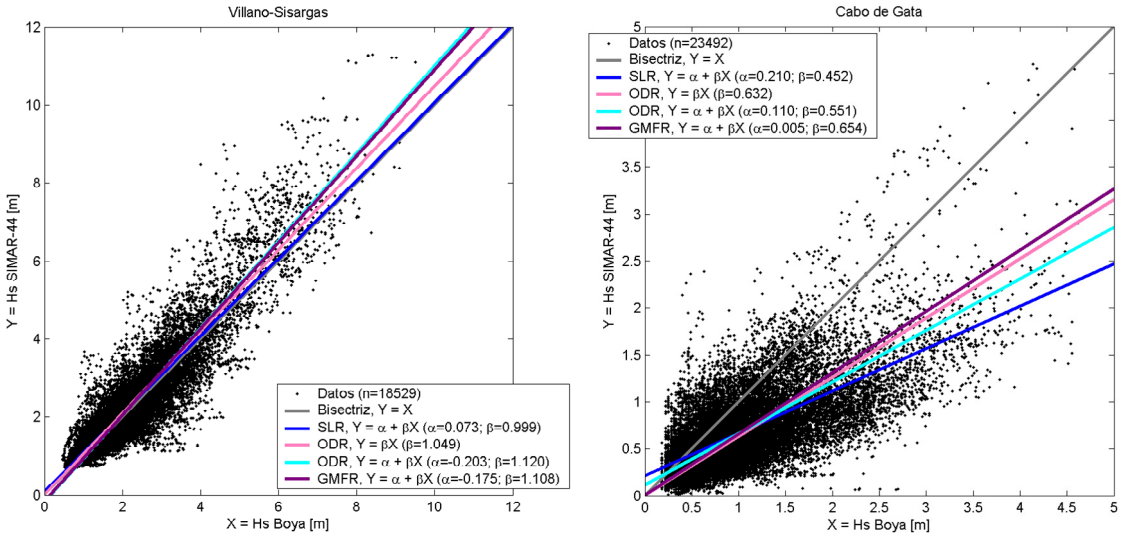

Figura 3.14. Ajuste de distintas regresiones EIV y recta SLR a datos de H_s de dos boyas y sus correspondientes de SIMAR-44.

3.2.3.6. Aplicación de regresiones con tres variables (FR).

Una vez expuestos ejemplos de aplicación de regresiones bidimensionales, en este último apartado de regresiones se van a aplicar las regresiones FR (EIV múltiples) a dos casos tridimensionales (con tres variables).

Los datos utilizados son la H_s proveniente de tres bases de datos para dos ubicaciones distintas. La primera base de datos es el reanálisis de SIMAR-44 perteneciente al OPPE, que será la variable X, otra base de datos son las boyas de la red Exterior del OPPE, que será la variable Y y la otra base de datos son los datos de satélites pertenecientes al IHCantabria (GEOSAT, TOPEX/POSEIDON, GFO, JASON-1 y ENVISAT) que será la variable Z. Las dos ubicaciones donde se realizan las comparaciones son la posición de la boya de Estaca de Bares (Atlántico) y la de Mahón (Mediterráneo), en la tabla 2.1 y en la figura 2.15 del capítulo 2 se pueden ver dichas posiciones. Cabe señalar que son escasos los registros para los que se da la coincidencia espacial y temporal de las tres bases de datos, por eso se han seleccionado estas dos boyas, pues las trazas de los satélites son bastante próximas a sus posiciones de fondeo.

En la figura 3.15 se presentan los tres tipos de regresiones FR explicadas en el apartado 3.2.3.4, volviendo a representar el modelo de regresión de cada uno. Así se denota por FR1 (azul) el modelo de regresión con dos rectas, FR2 (verde) el modelo de regresión con dos

rectas que pasan por el origen y FR3 (rojo) el modelo de regresión con tres rectas que pasan por el origen.

$$\text{Regresiones FR (3 variables)}\begin{cases}\underline{\text{2 rectas (FR1)}}\begin{cases}X = T + e_X\\ Y = \alpha_1 + \beta_1 T + e_Y\\ Z = \alpha_2 + \beta_2 T + e_Z\end{cases}\\ \underline{\text{2 rectas por el origen (FR2)}}\begin{cases}X = T + e_X\\ Y = \beta_1 T + e_Y\\ Z = \beta_2 T + e_Z\end{cases}\\ \underline{\text{3 rectas por el origen (FR3)}}\begin{cases}X = \beta_1 T + e_X\\ Y = \beta_2 T + e_Y\\ Z = \beta_3 T + e_Z\end{cases}\\ \ldots\end{cases}$$

Figura 3.15. Tipos de regresiones FR con tres variables.

Cabe señalar que además de para calibrar bases de datos, estas regresiones pueden utilizarse también para homogeneizar distintas variables. Por ejemplo, tomando X como referencia, se pueden determinar las expresiones que relacionan Y y Z, gracias a la naturaleza simétrica de este tipo de regresiones. Así, siguiendo dichas expresiones (ver apartado 3.2.3.4), se pueden estimar los parámetros que definen las relaciones entre X, Y y Z (SIMAR-44, Boyas y Satélites) para las H_s de las posiciones de las boyas de Estaca de Bares y Mahón. A modo de ejemplo, en la figura 3.16 se presentan únicamente las expresiones que ligan Z con Y para las tres regresiones FR, con dichas relaciones se puede homogeneizar los datos de los satélites y los de las boyas.

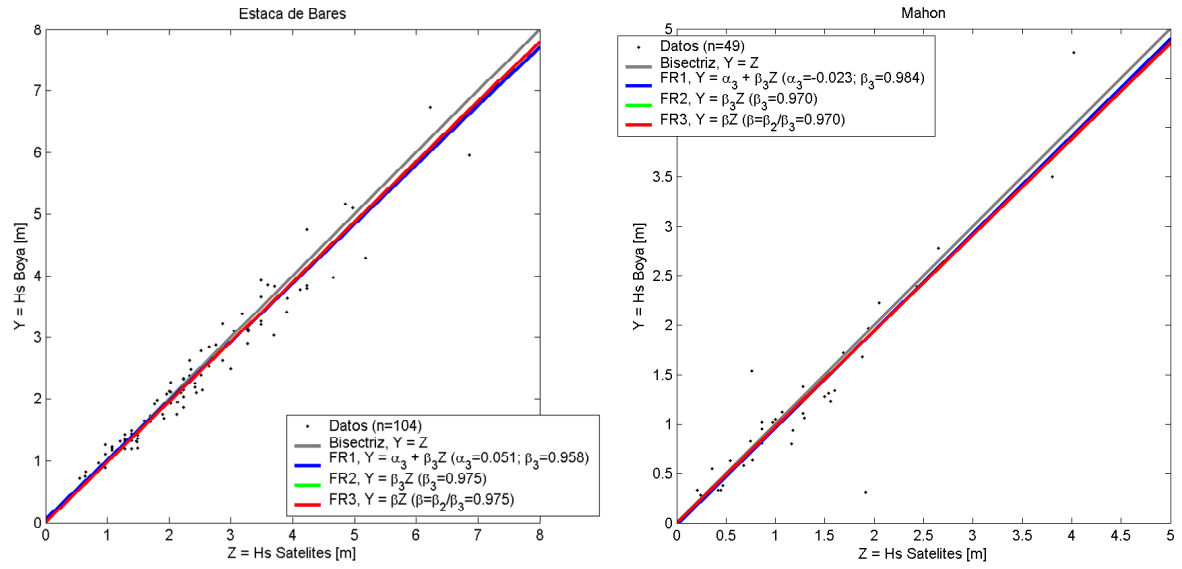

Figura 3.16. Ajuste de distintas regresiones FR a los datos comunes de H_s de dos boyas, de Satélites y SIMAR-44. Se representan únicamente las relaciones de satélites frente a boyas.

Finalmente, es necesario reseñar que estas regresiones de la figura 3.16 han sido estimadas a partir de una población de datos muy escasa, pues es necesario que los datos de las tres fuentes sean coincidentes en tiempo (se han acotado sus diferencias a 30 minutos) y en posición (para el Atlántico se ha permitido diferencias de 0.5º y en el Mediterráneo de 0.25º). Es por ello que se tienen tan pocos datos coincidentes anteriores al 2001 (fin de los datos de SIMAR-44) con estos criterios, aproximadamente entre 50 y 100 datos. Debido a ello las expresiones pueden no ser muy significativas o generalizables, lo que sí se puede comprobar es que las tres relaciones en ambas posiciones son muy parecidas y próximas a la recta bisectriz, pudiendo verificarse que, en estos casos, boyas y satélites miden aproximadamente con las mismas incertidumbres, de hecho las expresiones FR2 y FR3 son exactamente iguales. Eso es debido a que el método iterativo de resolución de FR3 identifica X con T de manera que resulta más sencillo, para este caso, la utilización de las expresiones explicitas de FR2.

3.2.4. Diagnóstico.

En el apartado anterior se han presentado distintas regresiones que ligan dos variables que se pueden utilizar para comparar los distintos modelos de regresión y evaluar el óptimo. En este epígrafe se va a describir algunas técnicas para diagnosticar una variable o ajuste en función de la variable original o de referencia. De esta manera, se obtendrán conclusiones sobre una determinada relación de calibración o ajuste a un modelo, o los errores cometidos por las aproximaciones efectuadas.

Se va a evaluar la semejanza entre dos variables, es decir se van a definir una serie de argumentos para cuantificar esa similitud. Por ejemplo, anteriormente cuando se ha evaluado las regresiones en dos casos concretos (Villano-Sisargas y Cabo de Gata) se han utilizado términos como “tiene una mayor dispersión” o “tiene una menor correlación” sin cuantificar dichos valores. Con las parametrizaciones siguientes se podrá valorar de manera no subjetiva las relaciones entre dos variables.

Estas técnicas de comparación o diagnóstico se pueden basar en la comparación dato a dato, propiamente dichos, o en la comparación de los regímenes medios de dichas variables. A continuación se presentarán ambas aproximaciones, que dan informaciones complementarias, aplicándolas a los casos anteriormente expuestos, la H_s proveniente de dos bases de datos (SIMAR-44 y la red de boyas exteriores de OPPE) para dos ubicaciones distintas (Villano-Sisargas y Cabo de Gata).

3.2.4.1. Diagnóstico de los datos.

La primera forma de diagnosticar dos variables es mediante la comparación directa de los datos de que se dispone. La representación mediante punteo de un dato frente a su dato coincidente se denomina generalmente diagrama de dispersión (*scatter plot*), a partir del cual se han realizado los ajustes de las distintas regresiones. El diagnostico de los datos se va a realizar definiendo una serie de parámetros ligados a la regresión clásica, pues es la más sencilla y es la que se sigue en la literatura, pues se considera que el error lo tiene la variable a evaluar (Y o respuesta) y no la de referencia (X o covariable).

No se va a entrar en esta tesis en la eliminación de datos fuera de tendencia o *outliers*, pues se considera que los datos almacenados en las bases de datos han pasado unos controles de calidad que los eliminan; de todas formas si se quiere más información sobres las técnicas de eliminación de *outliers* se puede consultar por ejemplo Barnett y Lewis (1994).

Para el diagnostico de los datos, si se representa mediante punteado X (datos de referencia: $x_1, x_2, ..., x_i, ..., x_n$) frente a Y (datos a evaluar: $y_1, y_2, ..., y_i, ..., y_n$), definiendo cada par de datos un punto (x_i, y_i), lo óptimo sería que coincidiesen todos los puntos en la bisectriz (recta $Y = X$). Para cuantificar la bondad de ajuste, a lo largo de toda esta tesis y de manera recurrente, se van a definir una serie de parámetros ($BIAS$, RMS, ρ y SI) que cuantifican la desviación de dichos puntos respecto de la recta bisectriz. También existen otros parámetros o representaciones, como los diagramas de Taylor (Taylor, 2001), que expresan gráficamente la información de $BIAS$, RMS y ρ, pero por su sencillez y generalización se van a utilizar los cuatro parámetros definidos a continuación:

- *BIAS*

El sesgo o *bias* es la desviación sistemática entre dos variables y se define como:

$$BIAS = \overline{x} - \overline{y} \tag{3.57}$$

Mide la diferencia de las medias de ambas variables, dando información de cuanto difieren entre sí los momentos de orden 1.

- *RMS*

El error cuadrático medio (RMS, *Root Mean Square*) mide la exactitud con que se parecen dos variables, teniendo en cuenta tanto el cuadrado del sesgo como la varianza o precisión entre ellas; se define como:

$$RMS = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2}$$

(3.58)

Se define como la raíz cuadrada de la esperanza del cuadrado de la diferencia entre dos variables y contiene información de los momentos de orden 1 y 2.

- ρ

El coeficiente de correlación de la regresión, para el modelo de regresión $\hat{y}_i = x_i$ (recta bisectriz), se denomina coeficiente de correlación de la recta bisectriz. Mide la intensidad de la relación de igualdad entre dos variables. Está definido entre 0 y 1, cuando existe correlación perfecta entre las dos variables (los datos de X e Y son iguales) $\rho = 1$.

$$\rho = \sqrt{R^2}$$

(3.59)

Siendo R^2, para este caso, el porcentaje de ajuste, en tanto por uno, entre las dos variables (o debida a la recta bisectriz):

$$R^2 = \frac{SSR}{SST} = \frac{SSR}{SSE + SSR} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2}{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + (\hat{y}_i - \overline{y})^2} = \frac{\sum_{i=1}^{n}(x_i - \overline{y})^2}{\sum_{i=1}^{n}(y_i - x_i)^2 + (x_i - \overline{y})^2}$$

(3.60)

- SI

El índice de dispersión (SI, *Residual Scatter Index*) respecto a la regresión recta bisectriz, se define como:

$$SI = \frac{\sqrt{s_E^{\,2}}}{\overline{x}} = \frac{\sqrt{\frac{SSE}{n}}}{\overline{x}} = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}}{\overline{x}} = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - x_i)^2}}{\overline{x}} = \frac{RMS}{\overline{x}}$$

(3.61)

Mide la dispersión de los puntos respecto a la recta bisectriz; si todos los puntos se sitúan sobre la bisectriz, el parámetro adimensional SI toma el valor 0.

En la figura 3.17 se presentan a la misma escala los diagramas de dispersión de H_s del caso de Villano-Sisargas y Cabo de Gata, calculando los cuatro parámetros de diagnóstico de los datos ($BIAS$, RMS, ρ y SI), verificándose que los datos de SIMAR-44 se parecen más a los de la boya de Villano-Sisargas que en la posición de Cabo de Gata. El sesgo o $BIAS$ es casi 0 en Villano-Sisargas, a pesar de los grandes valores de H_s; en cambio Cabo de Gata,

con valores mucho menores de H_s tiene casi 35 cm. Al igual que la diferencia de momentos de orden 1 ($BIAS$) es mayor para Cabo de Gata, su RMS también es superior al de Villano-Sisargas (59 cm frente a 56 cm), a pesar de que el oleaje es mucho menos energético en Cabo de Gata. Los otros dos parámetros presentan valores adimensionales respecto de la recta bisectriz también mejores para Villano-Sisargas, con un coeficiente de correlación (ρ) de 0.901 (más del 90%) y un índice de dispersión (SI) de 0.25, frente al 0.784 (78.4%) y 0.58 respectivamente de Cabo de Gata.

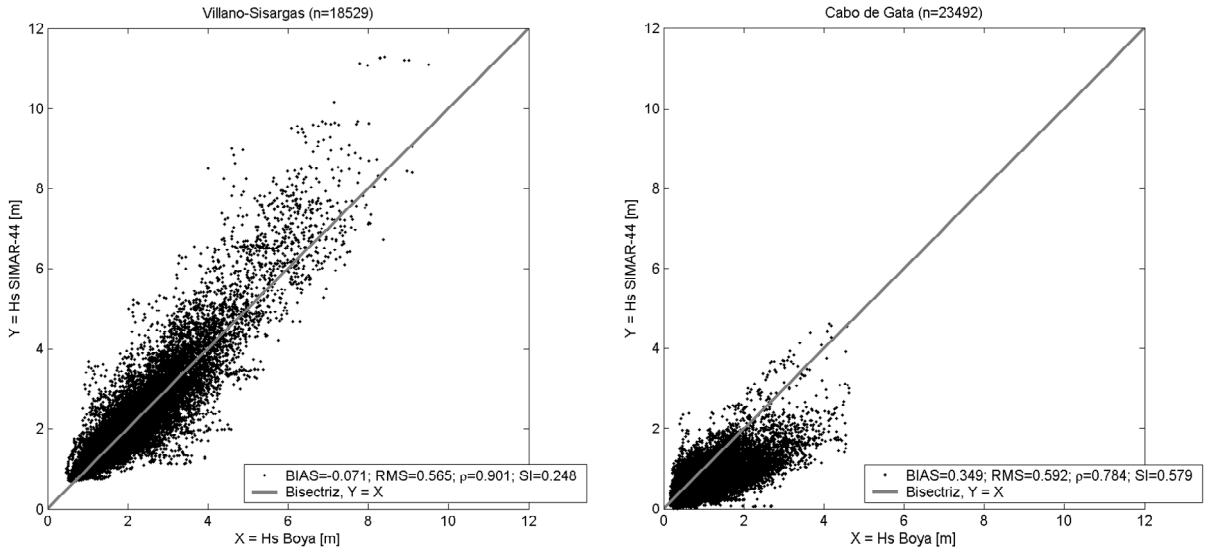

Figura 3.17. Diagnóstico de H_s de SIMAR-44 con datos de dos boyas de la red exterior de OPPE.

3.2.4.2. Diagnóstico del régimen medio.

Con los cuatro parámetros presentados en el diagnostico de los datos se puede cuantificar la similitud entre dos variables, analizando tanto la tendencia central de los datos como su dispersión; pero dichos parámetros engloban las propiedades de todos los datos. Para poder ver las características específicas de los distintos valores que toma cada variable o los diferentes intervalos de probabilidad se utilizan métodos gráficos que comparan los regímenes medios de los datos. Estos gráficos no permiten evaluar la bondad de ajuste de forma cuantitativa, pero aportan una valiosa información cualitativa sobre el ajuste del régimen medio de los datos a evaluar respecto de los de referencia.

Básicamente existen dos métodos gráficos, los gráficos PP (Probabilidad-Probabilidad, *PP-plot*) y los gráficos QQ (Cuantil[5]-Cuantil, *QQ-plot*). Los distintos tipos y sus variantes de gráficos representan la misma información, pero a distinta escala. A continuación se van a ir

[5] Cuantil de orden a (x_a): Es el valor de la variable aleatoria X cuya probabilidad de no excedencia es a, por lo que se cumple $a = F(x_a)$. Por ejemplo, la mediana es el cuantil de orden 0.5.

presentando distintas alternativas gráficas para comparar la H_s proveniente de dos bases de datos (SIMAR-44 y boyas de la red exterior de OPPE) para las mismas dos ubicaciones presentadas anteriormente (Villano-Sisargas y Cabo de Gata).

Los gráficos PP representan, mediante punteado, la probabilidad de no excedencia que toman cada una de las dos variables para un mismo conjunto de valores. En la figura 3.18 se puntean las probabilidades que toman los datos de la boya ($F(x)$) y las de SIMAR-44 ($F(y)$) para un mismo conjunto de 30 valores de H_s equiespaciados en altura de ola desde el valor mínimo de H_s de la boya hasta su valor máximo. Se puede comprobar como los puntos en Villano-Sisargas se sitúan sobre la bisectriz (mostrando para los valores evaluados que las dos funciones de distribución coinciden). En cambio, para Cabo de Gata las probabilidades medias difieren mucho de la bisectriz, reflejando como en gráficos anteriores que los datos de SIMAR-44 son bastante distintos a los de la boya exterior de Cabo de Gata. La distribución de valores de H_s elegidos para calcular sus probabilidades, equiespaciada en H_s, hace que en los gráficos PP aparezcan muchos puntos en la parte alta del régimen medio (próximos a 1).

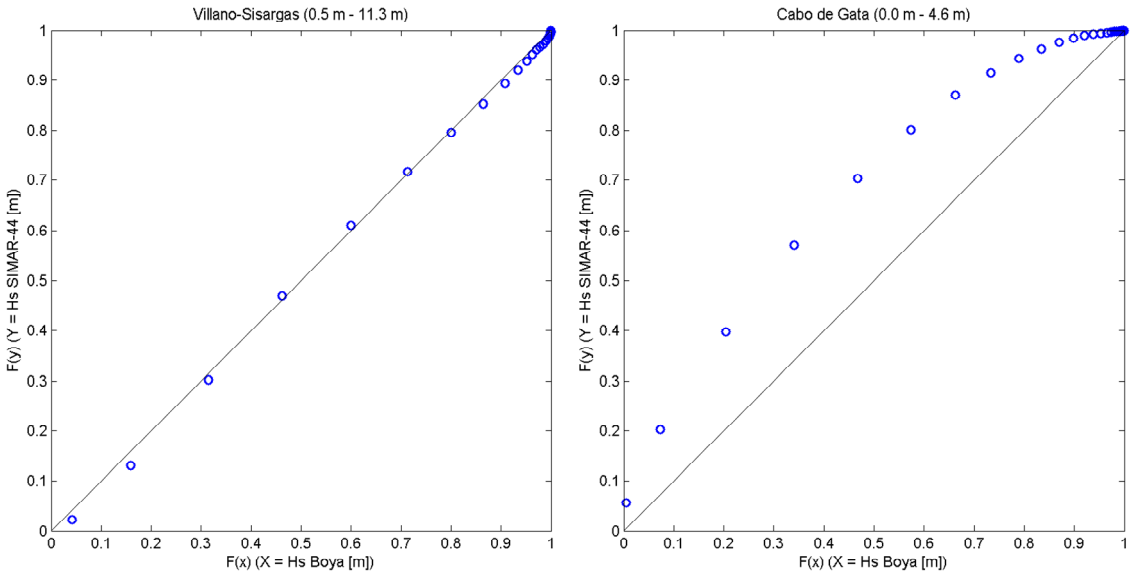

Figura 3.18. Gráficos PP tomando 30 datos de H_s equiespaciados en H_s (desde el mínimo al máximo de las boyas).

Para evitar que en los gráficos PP aparezcan demasiados puntos próximos a la probabilidad 1, en la figura 3.19 se presenta otro tipo de gráfico PP que no toma 30 datos de H_s equiespaciados (como en la figura 3.18). Así en esta figura se eligen 30 valores de H_s cuyas probabilidades de no excedencia en la variable X (boya de la red exterior) están equiespaciadas (desde el 1% al 99%), o lo que es lo mismo se toman los cuantiles de X para las 30 probabilidades equiespaciadas. A partir de esos 30 valores de H_s se calculan las probabilidades en la variable Y (SIMAR-44) y se puntean frente a las de X (las

probabilidades originalmente equiespaciadas desde el 1% al 99%). Se puede comprobar como ambos gráficos PP muestran la misma información, los puntos en Villano-Sisargas se sitúan sobre la bisectriz, no siendo así en Cabo de Gata para las probabilidades medias.

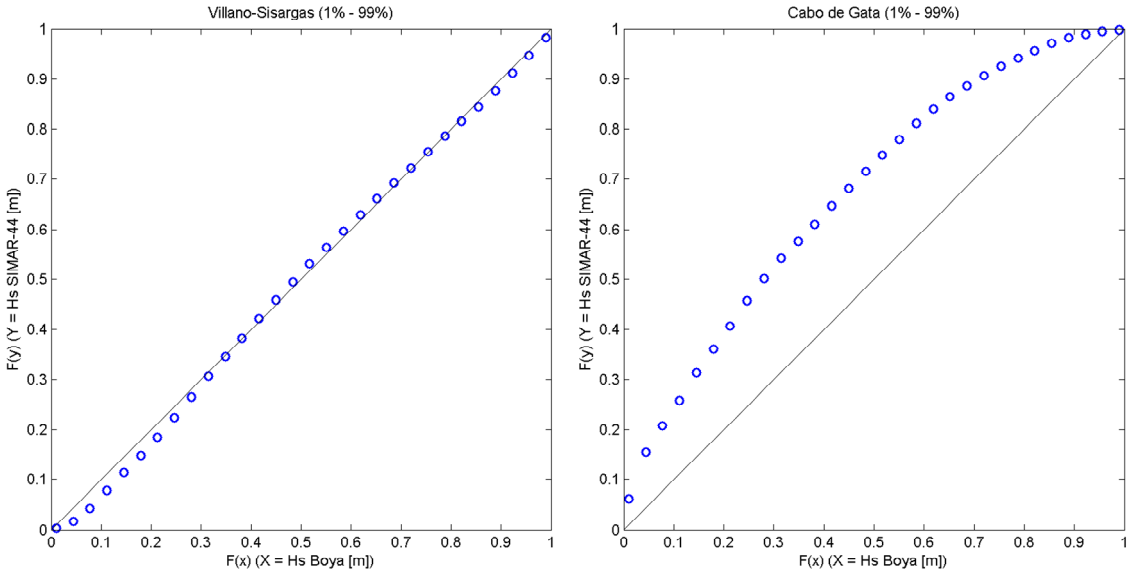

Figura 3.19. Gráficos PP tomando 30 datos de cuantiles de H_s de las boyas equiespaciados en probabilidad (del 1% al 99%).

Cabe señalar que los gráficos PP siempre tienden a pasar por 0 y 1 pues los valores mínimos y máximos de las distintas variables, sean los que sean, toman valores de probabilidad próximos a 0 y 1 respectivamente. Por ello con los gráficos PP no se permite distinguir si la parte baja y alta del régimen medio se parecen o no.

Los gráficos QQ permiten verificar si las distintas partes del régimen medio de dos variables se parecen. Estos métodos gráficos se basan en la representación, mediante punteado, de distintos cuantiles obtenidos de los dos regímenes (o funciones de distribución) a comparar. Por ejemplo en la figura 3.20 se representan los *QQ-plot* tomando los cuantiles (de ambas variables X e Y) para las 30 probabilidades equiespaciadas desde el 1% al 99%, los mismos que en la figura 3.19. Se puede verificar que, como en los gráficos PP, las probabilidades medias de SIMAR-44 en Cabo de Gata difieren de la boya, no siendo así para Villano-Sisarga; pero ahora también se puede comprobar que la rama alta del régimen medio de SIMAR-44 es mayor que la de la boya de Villano-Sisarga, siendo la de Cabo de Gata mucho menor.

Como puede observarse en éstas gráficas, los *QQ-plot* están escalados en función de los valores que toma cada variable (Villano-Sisargas tiene $0\ \text{m} \le H_s \le 12\ \text{m}$ y Cabo de Gata $0\ \text{m} \le H_s \le 5\ \text{m}$), por el contrario los *PP-plot* siempre toman valores entre 0 y 1.

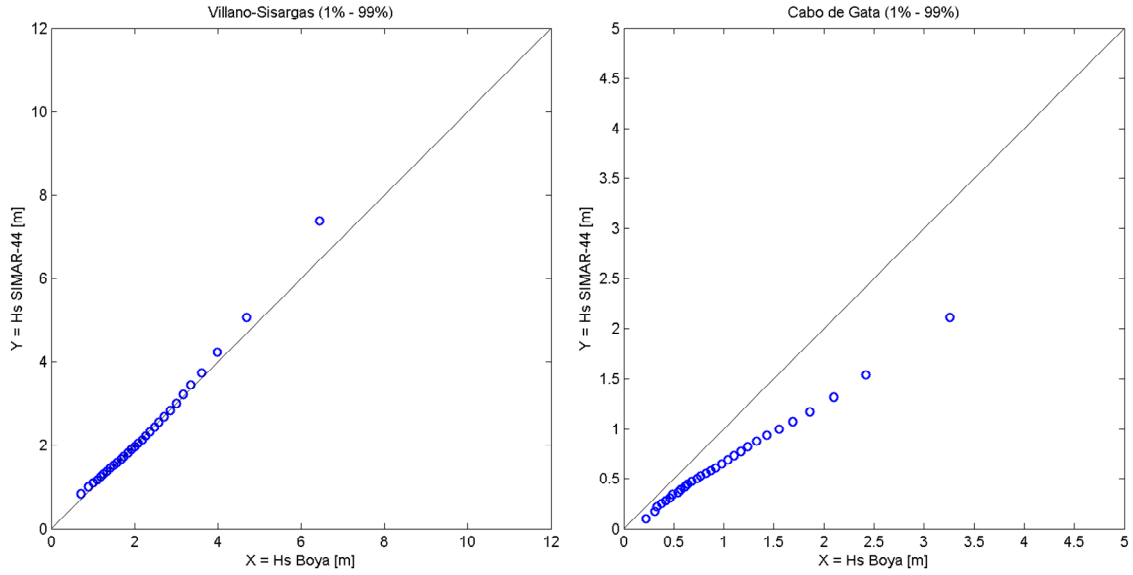

Figura 3.20. Gráficos QQ tomando 30 datos de cuantiles de H_s equiespaciados en probabilidad (del 1% al 99%).

Existen muchas opciones para determinar los cuantiles de comparación para los gráficos QQ; por ejemplo equiespaciados en probabilidad (como el presentado en la figura 3.20 y similar a la figura 3.19, *PP-plot*); también podrían ser cuantiles cuyas H_s estén equiespaciadas (análogo a la figura 3.18, *PP-plot*). En esta tesis la forma de seleccionar las probabilidades para determinar los cuantiles de comparación es equiespaciando la transformación de la probabilidad, $-\log[-\log(\Pr)]$, que es la variable reducida en el papel probabilistico de Gumbel de máximos (ver anejo I). Es decir, se seleccionan una serie de probabilidades cuyas variables reducidas en papel de Gumbel queden equiespaciadas.

En la figura 3.21 se presenta un ejemplo de este último tipo de gráficos QQ (con cuantiles equiespaciados en $-\log[-\log(\Pr)]$), que definen con aproximadamente la misma cantidad de cuantiles de comparación todo el intervalo de valores de H_s. Se puede comprobar como muestra las mismas tendencias para los valores medios y altos que el otro gráfico QQ (figura 3.20), si bien ahora la parte más extrema del régimen medio puede compararse con más nitidez, difiriendo cada vez más en Villano-Sisargas y cada vez menos en Cabo de Gata. La elección de los 30 cuantiles muestra el interés de la presente tesis por los mayores valores del régimen medio, pues se llega hasta probabilidades mayores que en el resto de casos, siendo la máxima probabilidad utilizada la que verifica $\Pr < 1 - 5/n$ [6].

[6] La relación $\Pr < 1 - 5/n$ determina la máxima probabilidad de no excedencia hasta la cual se pueden definir los intervalos de confianza de los cuantiles mediante las técnicas estadísticas clásicas (ver por ejemplo Luceño, 1989).

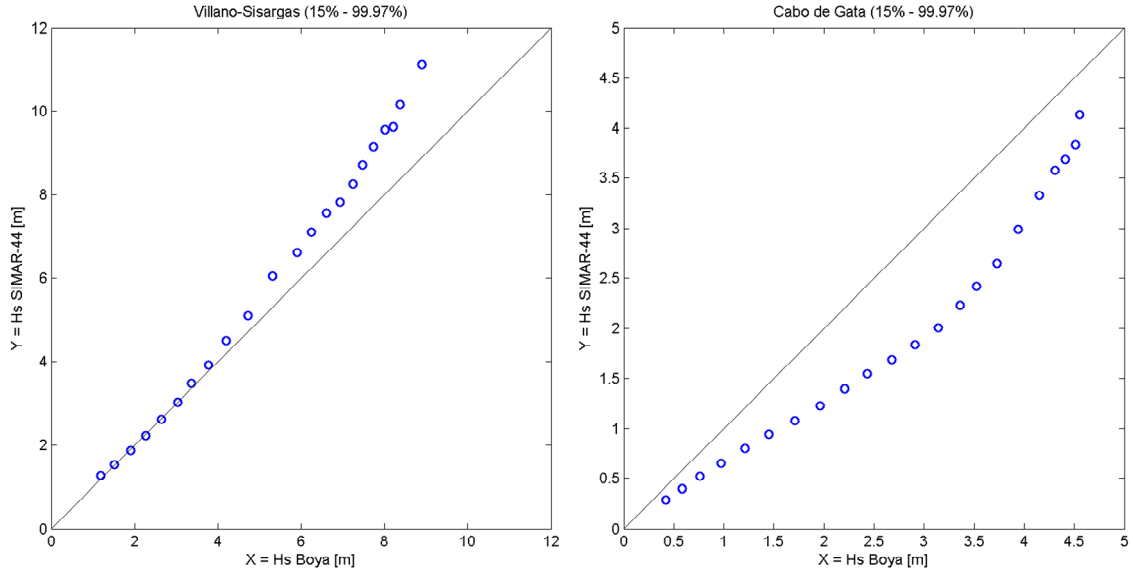

Figura 3.21. Gráficos QQ tomando 30 datos de cuantiles de H_s de las boyas equiespaciados en $-\log[-\log(\Pr)]$.

3.3. Estado del arte de las metodologías de calibración.

Una vez estudiadas y caracterizadas las distintas bases de datos de oleaje y posteriormente descritos los métodos o técnicas para modificar o tratar dichas bases de datos; en este apartado se va a presentar los métodos actuales para modificar dichas bases de datos mejorando su calidad (calibración), así como su evolución a lo largo del tiempo como resultado de la aparición de nuevas herramientas para medir o representar esa realidad. Por lo tanto, este estado del arte de las metodologías de calibración se va a presentar cronológicamente en función de las creación de las distintas bases de datos, visuales, satélites y finalmente de modelado numérico.

3.3.1. Datos visuales.

Históricamente, la primera fuente de información de datos de oleaje fue la visual (*Voluntary Observing Ships,* VOS). Desde mediados del siglo XIX y hasta la actualidad observadores entrenados desde las cubiertas de los barcos en ruta estiman información del oleaje visualmente. Muchos autores han estudiado las incertidumbres en las observaciones visuales (por ejemplo Cartwright, 1964; Hogben y Lumb, 1967; Nordenstrom, 1969; Wilkerson y Earle, 1990; Gulev y Hasse, 1998; Gulev *et al.*, 2003) pero a pesar de que dicha base de datos es muy dependiente del entrenamiento del observador, es la que tiene registros más antiguos y contiene información de la altura de ola, periodo y dirección tanto de la componente *sea* (mar de viento) como de la *swell* (mar de fondo) del oleaje.

Desde finales del siglo XIX existen distintos instrumentos mecánicos capaces de medir el oleaje, pero no es hasta principios del siglo XX cuando se utilizan aviones, barcos o boyas oceanográficas destinadas a medir el oleaje con instrumentos electrónicos. A partir de los años 40 entran en funcionamiento una serie de barcos meteorológicos (*Ocean Weather Station,* OWS) para navegar de manera continua por posiciones fijas que recogen información instrumental y visual del oleaje.

Con la información de dichos barcos meteorológicos se realizaron las primeras calibraciones o relaciones entre los datos visuales de OWS y parámetros de oleaje provenientes de los instrumentos de medida de los OWS, como H_s, T_m y T_p. En la tabla 3.2 se presentan múltiples relaciones lineales (en su mayoría recopiladas de Massel, 1996) que mayoritariamente utilizan la regresión lineal clásica (SLR, *Simple Linear Regression*, ver epígrafe 3.2.3.1.1), pues usualmente se ha considerado despreciable el error en las mediciones instrumentales (del orden de los centímetros en superficie libre) frente a las de los observadores (del orden de los decímetros). También existen relaciones ajustadas mediante la regresión lineal clásica que no son líneas rectas, que ligan la H_s y la altura de ola visual de los OWS, como la de Jardine (1979), $H_{OWS} = 0.22H_s^{\,2} + 0.78H_s + 0.83$ (unidades en m).

	Autores	Relaciones
H_s	Cartwright (1964)	$H_s = 1.28 + 0.88H_{OWS}$
	Hogben y Lumb (1967)	$H_s = 1.23 + 0.88H_{OWS}$
	Nordenstrom (1969)	$H_s = 1.51 + 0.85H_{OWS}$
	Hogben (1970)	$H_s = 1.19 + 0.52H_{OWS}$
	Hoffman y Miles (1976)	$H_s = 1.25 + 0.89H_{OWS}$
	Hoffman y Walden (1977)	$H_s = 2.13 + 0.78H_{OWS}$
	Jardine (1979)	$H_s = -0.51 + 1.02H_{OWS}$
	Quayle y Changery (1982)	$H_s = -0.10 + 1.38H_{OWS}$
	Soares (1986a)	$H_s = 1.47 + 0.84H_{OWS}$
T_m	Cartwright (1964)	$T_m = 5.19 + 0.37T_{OWS}$
	Hogben y Lumb (1967)	$T_m = 4.70 + 0.32T_{OWS}$
	Soares (1986b)	$T_m = 5.61 + 0.31T_{OWS}$
T_p	Hogben y Lumb (1967)	$T_p = 4.10 + 0.76T_{OWS}$

Tabla 3.2. Rectas de calibración de datos visuales OWS con datos instrumentales OWS (H en m y T en s).

Con la información visual recogida por los oceanógrafos del OWS también se realizan correlaciones con la visual voluntaria de los barcos comerciales (VOS). En la tabla 3.3 se presentan algunas de esas regresiones de rectas entre datos de altura de ola y de periodo de oleaje. A partir de la aplicación conjunta de las relaciones de calibración de las tablas 3.2 y 3.3, se realizan las primeras calibraciones generales de la base de datos visuales; de todas ellas, las más utilizadas son las de Soares (1986a) y Soares (1986b).

	Autores	Relaciones
H	Hogben y Lumb (1967)	$H_{OWS} = 1.5 + 0.75H_{VOS}$
	Soares (1986a)	$H_{OWS} = 1.02 + 0.89H_{VOS}$
T	Soares (1986b)	$T_{OWS} = 6.58 + 0.2T_{VOS}$

Tabla 3.3. Rectas de calibración de datos visuales VOS con datos visuales OWS (H en m y T en s).

En la tabla 3.4 se presenta unas relaciones que ligan directamente los datos medidos por los barcos meteorológicos (H_s y T_m ,) con los datos visuales, VOS. Se presentan dichas relaciones de Nordenstrom (1969), pues son de los primeros modelos de regresión no lineales que utilizan la curva potencial del tipo $H_s = \beta H_{VOS}^{\ \gamma}$ para calibrar.

	Autor	Relaciones
H_s	Nordenstrom (1969)	$H_s = 1.68H_{VOS}^{\ 0.75}$
T_m		$T_m = 2.83T_{VOS}^{\ 0.44}$

Tabla 3.4. Relaciones de calibración no lineales de datos visuales VOS con datos instrumentales (H en m y T en s).

Las expresiones de calibración mostradas hasta ahora están basadas en las medidas instrumentales de los barcos meteorológicos, pero también existen relaciones de calibración de datos visuales a partir de otra fuente de información, las boyas oceanográficas. Hasta los años 70-80 no se generaliza el uso de boyas fondeadas en una misma ubicación, creándose a partir de entonces las primeras redes de boyas permanentes. Las bases de datos de boyas oceanográficas son mucho más precisas que las visuales, sirviendo para evaluar sus errores. A partir de la información suministrada por las boyas se han producido una multitud de relaciones lineales para calibrar los datos visuales de la zona de influencia de cada boya utilizada; por ejemplo, en la tabla 3.5 se presentan algunas de las primeras rectas de calibración de H_s válidas para el ámbito costero español.

	Autores	Relaciones	
H_s	AMF (1991)	$H_s = 0.54 + 0.59H_{VOS}$	
	GIOC (1993)	$H_s = 0.5 + 0.6H_{VOS}$	$H_{VOS} > 0.5$

Tabla 3.5. Rectas de calibración de datos visuales VOS con boyas (unidades en m).

Posteriormente, y en la línea de Nordenstrom (1969), el "Atlas de Inundación del Litoral Español", en adelante ATLAS, desarrollado por el GIOC (1999) propone una metodología de calibración con relaciones potenciales del tipo $H_s = \alpha + \beta H_{VOS}^{\ \gamma}$ para corregir los regímenes medios de oleaje de datos visuales mediante un procedimiento iterativo de propagación hasta la posición de las boyas y comparación hasta la convergencia de ambos regímenes de oleaje (boyas y visuales calibrados). Dicha metodología parte de los regimenes medios direccionales

visuales caracterizados en la aún vigente ROM 0.3-91 (1992), que no incluyen ninguna calibración.

En el ATLAS se presenta una metodología de calibración comparando regimenes medios con una relación potencial entre las variables. Anteriormente Ochi (1978) y Dacunha *et al.* (1984) utilizaron el régimen medio conjunto de H_s y T_m, ajustando distribuciones Lognormal (ver anejo I) con datos de boyas, para estimar el valor medio de T_m a partir del de H_{VOS}, con expresiones del tipo $\mu_{LN}(T_m) = \alpha + \beta\mu_{LN}(H_{VOS})$. Finalmente Gulev y Hasse (1998), ajustando regimenes medios de H_s y T_m de boyas y barcos meteorológicos, obtuvieron diferentes relaciones logarítmicas de calibración de $T_{m\ sea}$ y $T_{m\ swell}$ a partir del $T_{VOS\ sea}$, $T_{VOS\ swell}$, $H_{VOS\ sea}$ y $H_{VOS\ swell}$ del tipo $T_m = a\log(T_{VOS} + b) + c\log(H_{VOS})$.

En la actualidad existen métodos generales de calibración para determinar H_s y T_m a partir de datos visuales (Gulev *et al.*, 2003) recopilando formulaciones como las de Gulev y Hasse (1998). Dicha metodología de calibración ha sido validada con otras fuentes de información. Como se detallará en los siguientes apartados, a partir de los años 80 y 90 se han producido intercomparaciones de datos provenientes de cuatro fuentes distintas de información: visuales, boyas, satélites y numéricos. Cada una de estas cuatro fuentes de información introduce errores distintos, por lo que Gulev *et al.* (1998) compara de dos en dos las diferentes H_s con regresiones lineales asumiendo errores en las distintas variables (*Error in Variables*, EIV) utilizando el método ODR (*Orthogonal Distance Regression*, ver epígrafe 3.2.3.3.1), en lugar de las regresiones lineales clásicas que desprecian el error cometido por una de las variables.

Hasta el momento no se han presentado calibraciones o correcciones en las direcciones de oleaje visual. Una de las primeras aportaciones a este tema se deben a DelBalzo *et al.* (2003) que introduce la comparación de la dirección, entre datos visuales y datos de boyas, estudiando la incertidumbre en su determinación.

3.3.2. Datos de satélites.

Otra fuente de información de datos de oleaje, que se empezó a desarrollar desde los años 70, fue la proveniente de altímetros situados a bordo de satélites. Estos instrumentos miden la altura de ola, aproximándose muy bien al parámetro H_s, dando también información sobre el periodo, pudiendo correlacionarse de manera más o menos eficaz con parámetros como T_m o T_p. Otros instrumentos situados en satélites también pueden dar información sobre el espectro del oleaje o su dirección, pero aún están en vías de desarrollo, por lo que en este apartado se describen sólo los datos de satélite provenientes de altímetros.

A lo largo de la historia se han ido incorporando distintas misiones con altímetros a bordo (SKYLAB, GEOS-3, SEASAT, GEOSAT, TOPEX/POSEIDON, ERS-1, ERS-2, GFO, JASON-1, ENVISAT y JASON-2) que han ido mejorando la precisión con la que se correlaciona la variable H_s. A pesar de ello, es necesaria su calibración para corregir distintos errores detectados en las diferentes misiones. Así las primeras misiones tenían errores del orden del metro en H_s (0.5 m tras el calibrado de GEOS-3, Gower, 1979), posteriormente la misión TOPEX/POSEIDON presenta errores del orden de los 25 cm en H_s (Fu y Cazedane, 2001) y las actuales con errores del orden de los centímetros (se espera que el altímetro de JASON-2 tenga una precisión de 3 cm).

En la tabla 3.6 se han recopilado distintas relaciones de calibración de H_s para cada una de las diferentes misiones de satélites con altímetros. Las primeras calibraciones se realizaron mediante regresiones lineales clásicas (SLR), utilizándose directamente los datos de H_s medidos por boyas coincidentes (en tiempo y posición) con los de satélites, (Gower, 1996, Carter *et al.*, 1992 o Barstow *et al.*, 1998), o indistintamente realizando la media mensual de los datos para obtener las relaciones de calibración (Cotton y Carter, 1994 o Young, 1999).

Posteriormente se han calibrado los datos de satélite con métodos EIV con datos instrumentales coincidentes. Challenor y Cotton (2002) y Soukissian y Kechris (2007) utilizan GMFR (*Geometric Mean Functional Relationship*, ver epígrafe 3.2.3.3.2). Queffeulou y Cotton (2002), Ray y Beckley (2003) y Soukissian y Kechris (2007) emplean el método ODR. Cabe señalar que este método asume que los errores de las distintas variables son iguales, a diferencia del GMFR que los considera distintos. Con el método ODR se pueden buscar relaciones polinómicas como la de Queffeulou y Cotton (2002) para ERS-1 u otras no lineales, de manera más sencilla que con GMFR.

Distintos autores han comparado las relaciones obtenidas por los tres métodos de regresión más utilizados para calibrar datos de oleaje provenientes de satélites: SLR y los dos EIV: ODR y GMFR. Se encuentran diferencias entre la aplicación de SLR y EIV (Ray y Beckley, 2003 y Soukissian y Kechris, 2007), en cambio no son significativamente distintas las expresiones de calibración de ODR y GMFR (Challenor y Cotton, 2002, Ray y Beckley, 2003 y Soukissian y Kechris, 2007). Esto es debido a que los errores cometidos por las mediciones de boyas y altímetros son similares, comprobado inicialmente por Carter *et al.* (1992) y Cotton y Carter (1994), entre otros; y recientemente por Caires y Sterl (2003) mediante métodos de comparación con tres fuentes de datos (boyas, satélites y numéricos). Debido a ello, las expresiones debidas a SLR y EIV son distintas, así como las de ODR y GMFR son iguales para las aplicaciones con datos de altura de ola entre boyas y satélites.

	Autores	Relaciones	
GEOSAT	Carter *et al.* (1992)	$H_s = 1.13 H_{GEOSAT}$	
	Cotton *et al.* (1998)	$H_s = 0.089 + 1.06 H_{GEOSAT}$	
	Barstow *et al.* (1998)	$H_s = 0.50 + 0.85 H_{GEOSAT}$	$H_{GEOSAT} < 1.77$
		$H_s = 1.13 H_{GEOSAT}$	$H_{GEOSAT} > 1.77$
	Young (1999)	$H_s = -0.148 + 1.144 H_{GEOSAT}$	
	Challenor y Cotton (2002)	$H_s = 0.1037 + 1.0999 H_{GEOSAT}$	
TOPEX	Cotton y Carter (1994)	$H_s = -0.19 + 1.09 H_{TOPEX}$	
	Gower (1996)	$H_s = -0.03 + 1.075 H_{TOPEX}$	
	Cotton *et al.* (1998)	$H_s = -0.082 + 1.049 H_{TOPEX}$	
	Barstow *et al.* (1998)	$H_s = -0.165 + 1.10 H_{TOPEX}$	
	Young (1999)	$H_s = -0.079 + 1.067 H_{TOPEX}$	
	Challenor y Cotton (2002)	$H_s = 0.0942 + 1.0523 H_{TOPEX}$	
		$H_s = 0.0827 + 1.0335 H_{TOPEX}$	Side-B
	Queffeulou y Cotton (2002)	$H_s = -0.0888 + 1.0658 H_{TOPEX}$	
		$H_s = -0.0674 + 1.0376 H_{TOPEX}$	Side-B
	Ray y Beckley (2003)	$H_s = -0.070 + 1.046 H_{TOPEX}$	Side-B
	Soukissian y Kechris (2007)	$H_s = -0.04637 + 0.9969 H_{TOPEX}$	Side-B
ERS-1	Cotton y Carter (1994)	$H_s = 0.136 + 1.267 H_{ERS-1}$	
	Cotton *et al.* (1998)	$H_s = 0.333 + 1.126 H_{ERS-1}$	
	Barstow *et al.* (1998)	$H_s = 0.47 + 1.01 H_{ERS-1}$	$H_{ERS-1} < 1.95$
		$H_s = 1.25 H_{ERS-1}$	$H_{ERS-1} > 1.95$
	Young (1999)	$H_s = 0.04 + 1.243 H_{ERS-1}$	
	Challenor y Cotton (2002)	$H_s = 0.3355 + 1.1091 H_{ERS-1}$	$t < 03/1995$
		$H_s = 0.2149 + 1.1273 H_{ERS-1}$	$t > 03/1995$
	Queffeulou y Cotton (2002)	$H_s = 0.19 + 1.19 H_{ERS-1}$	$t < 03/1995$
		$H_s = 0.4610 + 0.8684 H_{ERS-1} + +0.0558 H_{ERS-1}^{\ 2} - 0.0035 H_{ERS-1}^{\ 3}$	$t > 03/1995$ $H_{ERS-1} < 2.5$
		$H_s = 0.1069 + 1.1276 H_{ERS-1}$	$t > 03/1995$ $H_{ERS-1} > 2.5$
ERS-2	Challenor y Cotton (2002)	$H_s = 0.035 + 1.061 H_{ERS-2}$	
	Queffeulou y Cotton (2002)	$H_s = 0.0454 + 1.0627 H_{ERS-2}$	
GFO	Queffeulou y Cotton (2002)	$H_s = -0.0808 + 1.0633 H_{GFO}$	
JASON-1	Ray y Beckley (2003)	$H_s = -0.104 + 1.100 H_{JASON-1}$	

Tabla 3.6.. Relaciones de calibración de H_s provenientes de satélites (unidades en m).

En la tabla 3.6 se han enumerado relaciones de calibración para H_s almacenadas en bases de datos (OPR, *Off-line PRoducts*) que han sido procesados mediante algoritmos que garantizan un buen control de calidad, pero su ejecución tarda un tiempo no despreciable. Para oceanografía operacional es muy importante el tiempo en que llega un determinado dato, por lo que misiones como ERS-1 o ERS-2 proporcionan datos casi en tiempo real con un procesado rápido que proporciona un control de calidad resumido. Éstos datos (*Fast Delivery*)

requieren relaciones de calibración distintas a las dadas en la citada tabla, ejemplos de éstas expresiones se pueden tomar de Cotton *et al.* (1997).

En resumen, de todas las expresiones que aparecen en la tabla 3.6, en función del método de calibración y la generalidad de las expresiones (debida a la mayor información disponible para calibrar) se suelen utilizar las expresiones de Queffeulou y Cotton (2002), o en su defecto las de Challenor y Cotton (2002) o Ray y Beckley (2003). No obstante, cabe señalar que en la actualidad los organismos que distribuyen los datos de satélites proporcionan los datos de H_s calibrados o con referencias e instrucciones detallando la forma de realizarlo.

Otro tipo de información del oleaje que se puede correlacionar a partir de datos de satélite es el periodo y como ya se ha comentado, existen dos aproximaciones para determinarlo (ver apartado 2.4.4 del capítulo 2). Hwang *et al.* (1998) propone relaciones de T_m y T_p a partir de la relación teórica de H_s y U_{10} (velocidad del viento a 10 m sobre la superficie) para espectros saturados tipo *sea*, utilizando la regresión ODR para rectas que pasan por el origen. Actualmente existen más expresiones que siguen a Davies *et al.* (1998), que relacionan T_m y T_p a partir del parámetro P (ver ecuación 2.14) considerando las regresiones EIV.

En la tabla 3.7 se presentan algunas de las relaciones más comúnmente empleadas que ligan el parámetro P con T_m y T_p. De ellas, Gommenginger *et al.* (2003) utiliza regresiones lineales y log−log determinando los parámetros de ajuste con el método ODR. Posteriormente, a partir de las expresiones lineales de Gommenginger *et al.* (2003) para T_m. Por otro lado, Caires *et al.* (2005) modifica la calibración para oleajes tipo *sea*, SR<0.9 (*Swell Ratio*, $\mathrm{SR} = H_{s\,swell} / H_s$). Para determinar esta nueva expresión, Caires *et al.* (2005) utilizan un método de regresión que contiene información de tres fuentes: boyas, satélites y modelado numérico, denominado FR, *Functional Relationship* (ver epígrafe 3.2.3.4).

	Autores	Relaciones	
T_m	Gommenginger *et al.* (2003)	$T_m = -0.895 + 2.545P$	
		$\log_{10} T_m = 0.361 + 0.967 \log_{10} P$	
	Caires *et al.* (2005)	$T_m = -0.895 + 2.545P$	SR>0.9
		$T_m = 0.97 + 1.78P$	SR<0.9
T_p	Gommenginger *et al.* (2003)	$\log_{10} T_p = 0.154 + 1.797 \log_{10} P$	

Tabla 3.7. Relaciones de calibración de periodos de oleaje provenientes de satélites (unidades en s).

También existen relaciones de calibración determinadas con redes neuronales que relacionan H_s y σ_0 (potencia del pulso reflejado, *Normalized Radar Cross Section*) con T_m o también H_s, U_{10} y σ de ambas bandas de frecuencia (K_u y C) con T_m. Estas expresiones son exponenciales y con parametrizaciones complicadas, véase en Quilfen *et al.* (2005) más detalles para su aplicación.

Sin embargo, los datos de periodo de ola aún no disponen de relaciones tan precisas y generales como las presentadas para H_s. Debido fundamentalmente a que H_s se estima directamente con los altímetros, en cambio T_m y T_p se correlacionan a partir de otros parámetros medidos por los satélites como H_s, U_{10} o σ_0. Así por ejemplo en la figura 3.22 se presenta una comparación entre T_m (parte izquierda de la figura) de la boya de Mahón y los obtenidos a partir de las formulaciones de Gommenginger *et al.* (2003), Caires *et al.* (2005) y Quilfen *et al.* (2005) de datos de satélites (GEOSAT, TOPEX/POSEIDON, GFO, JASON-1 y ENVISAT) que distan menos de 0.25º de la posición de la boya. En general, se observa para las cuatro expresiones (comparadas con la boya de Mahón) que subestiman los menores valores de T_m y sobrestiman los mayores, siendo la expresión de Caires *et al.* (2005), ver tabla 3.7, la que mejores ajustes presenta para este ejemplo. No obstante dichos resultados de T_m no llegan a ajustarse con tanta precisión como los obtenidos para H_s (parte derecha de la figura 3.22).

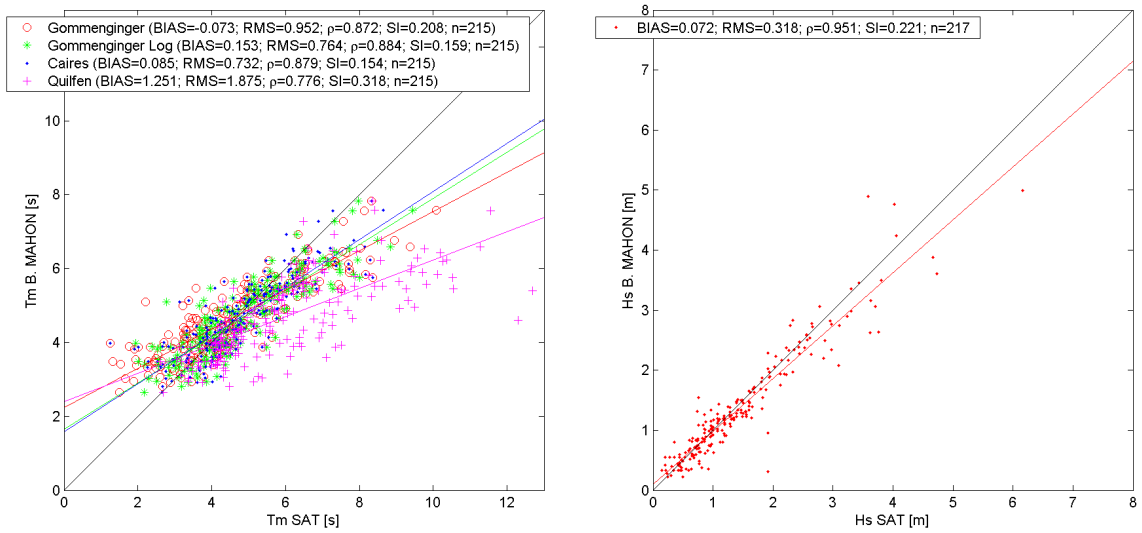

Figura 3.22. Comparación de varias relaciones de calibración de T_m de satélites con la boya de Mahón (izquierda) y comparación de H_s de satélites frente a la boya de Mahón (derecha).

3.3.3. Datos de modelado numérico.

Casi paralelamente a la aparición de los datos de oleaje provenientes de satélites, que dan una información global del oleaje, se desarrollaron modelos numéricos que son capaces de simular el oleaje también de forma global, resolviendo el espectro direccional del oleaje (con información tanto del *sea* como del *swell*).

Como es sabido, los datos obtenidos mediante modelado numérico o sintéticos no pueden utilizarse directamente (no son medidas de la naturaleza) ya que están sometidos a las

imprecisiones inherentes al cálculo matemático y a los forzamientos utilizados. Muchos y variadas formas de validar y calibrar estos datos han sido desarrollados a partir de las distintas formas de observar el oleaje (visuales, boyas, satélites, …). Una primera clasificación de los métodos de calibración consiste en separar los que calibran toda la base de datos (calibración global) y los que calibran la serie de oleaje en un sólo punto (calibración puntual).

No se va a entrar en explicar las diferentes técnicas de calibración puntual, pues son análogas a las desarrolladas para calibrar datos visuales o de satélites; en líneas generales hay dos clases: las basadas en los distintos tipos de regresiones y las basadas en comparar los regímenes de datos numéricos y datos instrumentales. Éste último es el más comúnmente utilizado y ha sido la práctica habitual de calibración en los proyectos del IHCantabria. Un ejemplo de aplicación que muestra el estado de arte actual en la calibración de oleaje de reanálisis con boyas mediante igualación de regímenes se puede ver en CEDEX (2008), que obtiene relaciones lineales y log-log de calibración.

Las calibraciones globales de bases de datos de reanálisis o provenientes de modelos numéricos han sido generalmente realizadas con métodos de regresión FR, con varias fuentes de información: boyas, satélites y modelado numérico. Ejemplos de estimación del error cometido por cada una de ellas, validando sus resultados, se puede encontrar en Janssen *et al.*, (2003) o en Caires y Sterl (2003), ver apartado 3.2.3.4. Un ejemplo de calibración de modelo numérico con datos de boyas y satélites, siguiendo el método de Janssen *et al.* (2003), se presenta en Cavaleri y Sclavo (2006) que calibra los resultados del modelo WAM del ECMWF para el Mediterráneo.

Hasta ahora, en el estado del arte de las metodologías de calibración, sólo se han presentado métodos de calibración paramétrica, es decir se busca una fórmula o relación analítica que liga una variable con su corrección o calibración. Dicha relación tiene una serie de parámetros que son estimados eficientemente mediante distintos métodos. Sin embargo, también existen técnicas que modifican o distorsionan una determinada variable, calibrándola, sin la necesidad de definir ninguna relación paramétrica de calibración.

Éste es el caso de calibración no paramétrica de H_s definida en Caires y Ferreira (2005) que ha sido aplicado para la calibración global de, entre otros, el reanálisis ERA-40 a partir de datos de la misión TOPEX/POSEIDON (Caires y Sterl, 2005). Dicho método se basa en el estudio de series o subseries temporales, de manera que para cada valor de la serie a calibrar se busca el conjunto de valores en esa misma serie cuya historia reciente sea similar; de entre todos esos valores se seleccionan los que tienen información instrumental coincidente (en este caso datos de satélite). Con el promedio de las diferencias entre los datos de la serie original

seleccionados y sus datos instrumentales simultáneos se corrige cada uno de los valores de la serie; calibrando uno a uno los datos del reanálisis. En el capítulo 3 se ahondará en las características de dicho método, aplicándolo con ciertas variaciones.

Todas las calibraciones presentadas hasta el momento corrigen las distintas variables del oleaje como escalares, sin tener en cuenta la dirección de procedencia del oleaje. Camus *et al.* (2007) presenta una calibración direccional de H_s, es decir, se agregan los datos por direcciones dando diferentes correcciones a cada una. Esta metodología de calibración utiliza datos de satélites para determinar regresiones lineales del tipo $H_s = \beta H_{TOPEX}$ para cada dirección, verificando que las distintas familias de oleajes, según zonas y direcciones, necesitan correcciones muy distintas. En esta tesis se va a profundizar en dicha metodología de calibración direccional.

3.4. Conclusiones y consideraciones.

Un resumen de las conclusiones a las que se ha llegado tras el estudio del estado de conocimiento de las técnicas estadísticas de tratamiento de datos se presenta esquemáticamente en los siguientes puntos:

- De entre las técnicas estadísticas de tratamiento de datos, que ajustan datos a un modelo, el método de los mínimos cuadrados se utiliza básicamente para ajustar modelos de regresión que comparan distintos conjuntos de datos. Los otros tres métodos se utilizan para ajustar funciones de distribución o regímenes y son el método de los momentos, el de máxima verosimilitud y el de los papeles probabilísticos (basado en el de mínimos cuadrados).

- Hay múltiples tipos de regresiones para relacionar distintas variables de oleaje, la clásica, la simétrica, la EIV (la ODR, la GMFR,…), la FR, … La clásica es la que más se utiliza en general, la simétrica se ha utilizado de manera muy concreta antes de la aparición de la EIV, que es la que actualmente tiende a emplearse con mayor asiduidad para casos con dos variables. La FR es la que se emplea en estos momentos para calibraciones y comparaciones de tres variables, cuando existen varias fuentes de información de oleaje simultáneas (numéricos, boyas y satélites).

- De las regresiones lineales utilizadas en la actualidad se prefiere la utilización de métodos EIV frente a SLR, y dentro de ellos, a falta de nociones previas, es preferible la utilización del GMFR. Sin perjuicio de lo anterior, hay casos en los que los distintos

métodos dan los mismos resultados, por lo que no es inadecuado utilizar el método más sencillo en determinadas circunstancias.

- Las técnicas estadísticas de diagnóstico de datos que comprueban cuantitativamente si dos muestras de datos se parecen o no, se concretan tradicionalmente con la definición de cuatro parámetros ($BIAS$, RMS, ρ y SI). Para diagnosticar la proximidad entre los regímenes o funciones de distribución se utilizan métodos gráficos de comparación, de entre los que se prefiere el *QQ-plot* para comparar cualitativamente los regímenes medios de oleaje.

Atendiendo al estado del arte de las metodologías de calibración de las distintas bases de datos de oleaje se han llegado a las siguientes conclusiones:

- Inicialmente las bases de datos visuales fueron calibradas con regresiones SLR a partir de datos de barcos meteorológicos. Posteriormente, con información registrada por boyas, surgieron otras relaciones de calibración, tanto lineales como potenciales, pero fundamentalmente mediante la comparación de sus regímenes medios de oleaje y de validez fundamentalmente local. Finalmente, se ha desarrollado una metodología general de calibración, basada en la determinación de relaciones no lineales tras comparar los regímenes medios escalares de los datos visuales con los provenientes de boyas y barcos meteorológicos (Gulev *et al.*, 2003).

- La calibración de los datos de satélite prácticamente siempre se ha realizado comparando dato a dato coincidente con registros de boyas. Sólo en ciertas ocasiones se han utilizados algunos datos de barcos meteorológicos o inicialmente cuando se comparaban las medias mensuales de los datos. Dichas relaciones de calibración para H_s se obtuvieron siempre con regresiones, SLR muy inicialmente y ODR de manera general posteriormente. Las distintas relaciones de calibración varían para cada satélite, pero actualmente cada uno tiene muy caracterizados los errores que comete. Para la determinación de otros parámetros como los del periodo del oleaje aún no se tienen establecidas relaciones de calibración generales y para la dirección media del oleaje aún se investiga la forma de determinarla con precisión.

- Existen multitud de alternativas que tratan de determinar el periodo del oleaje a partir de otros parámetros de los satélites, pero de momento son únicamente validas para zonas próximas a las boyas utilizadas para su calibración, pues son muy dependientes de las condiciones de oleaje del lugar de donde son los datos de referencia para calibrar (por ejemplo, no correlacionan bien el periodo del oleaje tipo *swell* cuando no

hay viento). Se han utilizado regresiones ODR lineales o logarítmicas, regresiones lineales FR o incluso se buscan complicadas relaciones exponenciales con redes neuronales, con resultados aún no tan precisos como los de H_s de altímetros.

- En la actualidad, con la aparición de las bases de datos de reanálisis, se están haciendo muchos esfuerzos para su correcta validación y calibración. Por un lado, de manera más modesta y sólo para uso local, se realizan calibraciones punto a punto o de los regímenes medios de oleaje con datos instrumentales (generalmente sólo una boya) de la zona de interés, buscando relaciones lineales, logarítmicas o potenciales. De manera global también se busca la calibración de toda una base de datos de modelo numérico empleando técnicas más sofisticadas. Hasta el momento las dos más utilizadas son las regresiones FR lineales con datos de satélites y boyas y la no paramétrica definida por Caires y Ferreira (2005), aunque ninguna de ellas aborda el problema de la calibración direccional. Únicamente Camus *et al.* (2007) plantea el problema de la calibración direccional de datos de reanálisis con datos de satélites.

Finalmente se resumen una serie de consideraciones generales sobre el estado actual de las metodologías de calibración de bases de datos de oleaje:

- Existen multitud de metodologías de calibración de datos de oleaje. Mayoritariamente se buscan relaciones paramétricas sencilla mediante la comparación de datos coincidentes en tiempo y posición, pero también se pueden buscar relaciones entre los datos para conseguir que los regímenes sean lo más parecidos posibles. Esta última aproximación no requiere que los datos sean coincidentes en el tiempo, aunque es muy aconsejable.

- Cuando se realizan comparaciones dato a dato (*scatter plot*) entre dos variables generalmente se utilizan las regresiones para calibrar (casi siempre con modelos lineales) con lo que se mejoran los estadísticos medios, pero no se tienen en cuenta las diferentes tendencias que pueden tener los valores extremos frente a los medios. Una forma de resolver ese problema es buscando relaciones potenciales de calibración de manera que los regímenes de dos variables converjan.

- Las calibraciones cuando se tienen varias bases de datos (múltiples) se realizan actualmente con regresiones FR con modelos de regresión lineales. Estas técnicas también se usan para comparar y determinar el error que se comete con cada una de ellas, pudiendo utilizarse para homogenizar las diferentes bases de datos de manera que sea admisible su uso de manera conjunta.

- Como ya se ha comentado, con la utilización de regresiones para obtener relaciones paramétricas de calibración a veces no se calibra correctamente el régimen extremal del oleaje. Existen métodos no paramétricos que adoptan la distorsión necesaria para cada dato (sin ceñirse a ninguna relación paramétrica de calibración), obteniendo buenas calibraciones del régimen medio. Con dichos métodos no paramétricos se calibra o se modifican sólo los datos originales cuando se tiene información para ello, por lo que se necesitan suficientes datos de referencia para calibrar y que sean simultáneos a los originales (en el capítulo 7 se explica con detalle dicho método). Debido a ello generalmente no se modifica el régimen extremal.

- Las distintas metodologías de calibración existentes, en su inmensa mayoría, han despreciado la información direccional del oleaje, calibrando únicamente las variables escalares del oleaje con independencia de su dirección, aplicando elaborados modelos de regresión. Pero con dichos métodos de calibración escalar se enmascaran o distorsionan de manera incorrecta el oleaje en zonas donde las distintas direcciones del oleaje tienen necesidades de calibración diferentes, pues los reanálisis no siempre simulan con la misma calidad los oleajes de los diferentes sectores direccionales.

- Todos los tipos de calibraciones mencionadas son calibraciones puntuales. Es decir, calibran una determinada ubicación con información de las posiciones próximas, o sino utilizan toda la información disponible y realizan calibraciones de manera general de las bases de datos. Las calibraciones globales, que calibran en cada posición con las relaciones estrictamente necesarias, hasta ahora se resuelven con calibraciones puntuales aplicadas a pequeños dominios, limitando su uso a zonas con suficiente información de referencia.

- No existen metodologías de calibración que utilicen datos instrumentales con regímenes de oleaje sensiblemente distintos a los de la zona a calibrar (calibración espacial); siempre se basan en información suficientemente cercana como para asumir que el clima marítimo es similar.

Por tanto, el estado del arte sugiere la necesidad de:

- Definir una metodología general de calibración de bases de datos de oleaje para definir lo más correctamente posible (con la información disponible) el clima marítimo en una ubicación dada, especificando criterios para elegir los mejores métodos de calibración adaptándose a los datos disponibles e información necesaria.

- Que las distintas metodologías de calibración se planteen de la forma más general posible, aunque se apliquen al oleaje del ámbito costero español.

- Asegurar la rigurosidad estadística de las diferentes técnicas de calibración, y a su vez, utilizar las herramientas más simples y seguras, sin pérdida de precisión, para facilitar su uso y difusión.

- Definir criterios y recomendaciones para validar y verificar los resultados de las calibraciones.

- Que las distintas técnicas o métodos de calibración corrijan correctamente los oleajes más energéticos, que son los que condicionan el diseño de las obras marítimas.

- Definir métodos de calibración direccionales para corregir adecuadamente los oleajes que presentan diferenciadas características direccionales.

- Desarrollar un método de calibración espacial para resolver el problema de caracterizar el clima marítimo en una posición donde no hay datos instrumentales próximos.