Análisis de datos de producción

Energía Solar Fotovoltaica

Oscar Perpiñán Lamigueiro

Universidad Politécnica de Madrid

Planteamiento

- La energía producida por un SFV durante un período puede ser estimada a partir de la radiación incidente y de las características técnicas del sistema.
- Teniendo en cuenta el carácter estocástico de la radiación solar, la estimación de la energía que producirá un SFV durante los próximos años es un ejercicio de predicción con incertidumbre asociada.
- El funcionamiento de un SFV puede ser analizado tomando esta estimación como referencia (**comparación modelo-medidas**).
- En el caso de centrales FV, se pueden detectar problemas de funcionamiento comparando diferentes partes de la central (**coherencia estadística**).

Variable aleatoria y proceso estocástico

- Una **variable aleatoria** es una función que asigna un único numero real a cada resultado de un espacio muestral en un experimento.
- Un **proceso estocástico** es una variable aleatoria que evoluciona a lo largo del **tiempo** (p.ej. la radiación).

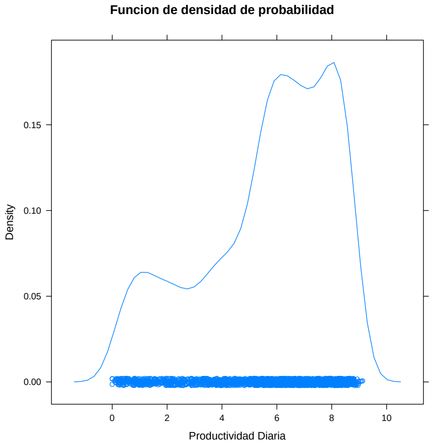
Función de densidad de probabilidad

La función de densidad de probabilidad, $f(X)$, de una variable aleatoria **asigna probabilidad** a un suceso:

$$P(a < X < b) = \int_a^b f(x)dx$$

$$P(X < b) = \int_{-\infty}^b f(x)dx$$

$$P(X > a) = \int_a^\infty f(x)dx$$



Media, varianza y desviación estándar

- La **media** de una variable aleatoria es el **centro de masas** de su función densidad de probabilidad:

$$\mu_X = \int_{-\infty}^{\infty} x \cdot f(x) dx$$

- La **varianza** de una variable aleatoria es la **media del cuadrado de las desviaciones** respecto a la media:

$$\sigma_X^2 = \int_{-\infty}^{\infty} (x - \mu_X)^2 \cdot f(x) dx$$

- La **desviación estándar** es la raiz cuadrada de la varianza: $\sigma_X = \sqrt{\sigma_X^2}$

Combinación lineal de variables aleatorias

- La **media de la suma** de varias variables aleatorias **independientes** es la suma de las medias:

$$\mu_{X_1+...+X_n} = \mu_{X_1} + ... + \mu_{X_n}$$

- La **varianza de la *suma o resta*** de varias variables aleatorias **independientes** es la **suma** de las varianzas:

$$\sigma^2_{X_1 \pm ... \pm X_n} = \sigma^2_{X_1} + ... + \sigma^2_{X_n}$$

Media y varianza de la media muestral

- Una **muestra de una población** es un conjunto de variables aleatorias independientes $(X_1 ... X_n)$.
- Si se toma una muestra de una población, la media de la muestra es otra variable aleatoria (que es una suma de variables aleatorias)

$$\overline{X} = \frac{1}{i = n} \sum_{n} X_i$$

Media y varianza de la media muestral

Sea una población cuya media es μ y su varianza es σ^2:

- La **media de la media muestral** es la media poblacional:

$$\mu_{\overline{X}} = \frac{1}{i = n} \sum_{n} \mu_{X_i} = \mu$$

- La **varianza de la media muestral** es la varianza poblacional dividido por el número de muestras.

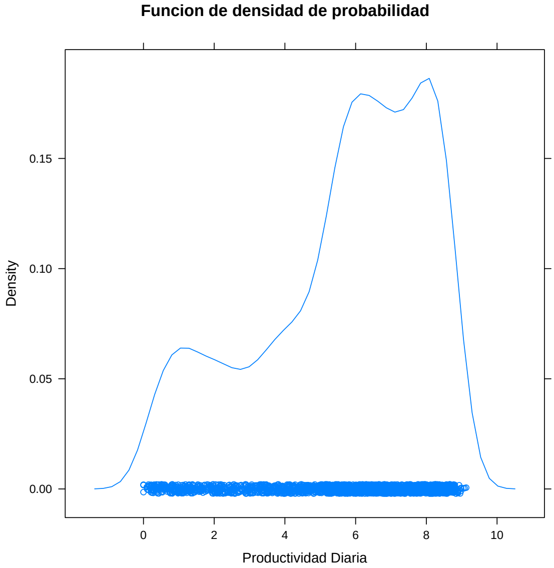
$$\sigma^2_{\overline{X}} = \sigma^2_{\frac{1}{n}X_1} + ... + \sigma^2_{\frac{1}{n}X_n} = \frac{\sigma^2}{N}$$

Por tanto, una forma de **reducir la incertidumbre** es realizar la **medida en repetidas ocasiones**.

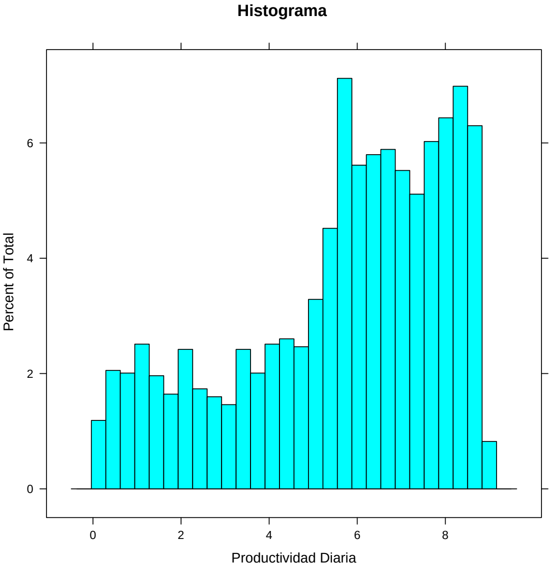
Mediana y cuartiles

- La **mediana** divide el conjunto de valores de la variable en **dos mitades** iguales (divide el area encerrada por la función densidad de probabilidad en dos partes iguales).
- Los **cuartiles** dividen este area en **cuatro** partes iguales.
- El area encerrada entre cada par de cuartiles es igual al 25% del total.
- La **mediana** es el **segundo cuartil**.
- La **distancia intercuartil** (definida entre los cuartiles 1 y 3) es una **medida de la dispersión** de la variable.

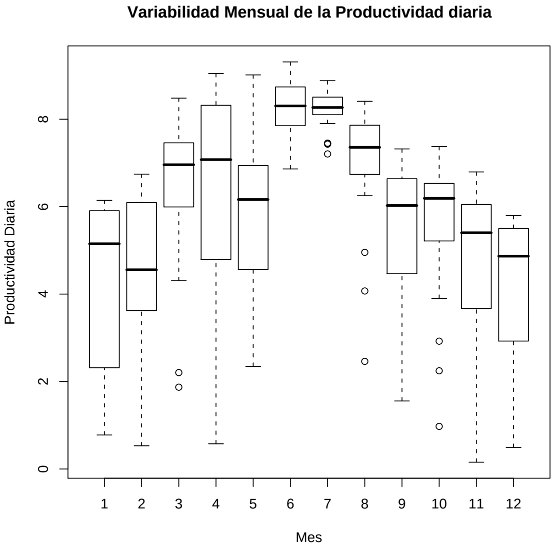
Función de Densidad de Probabilidad



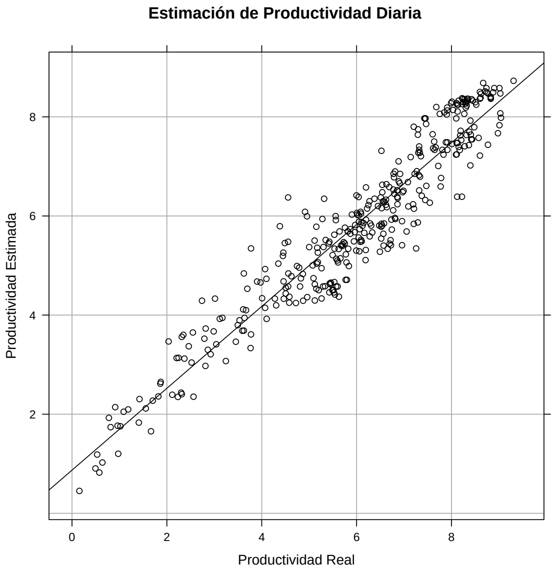
Histograma



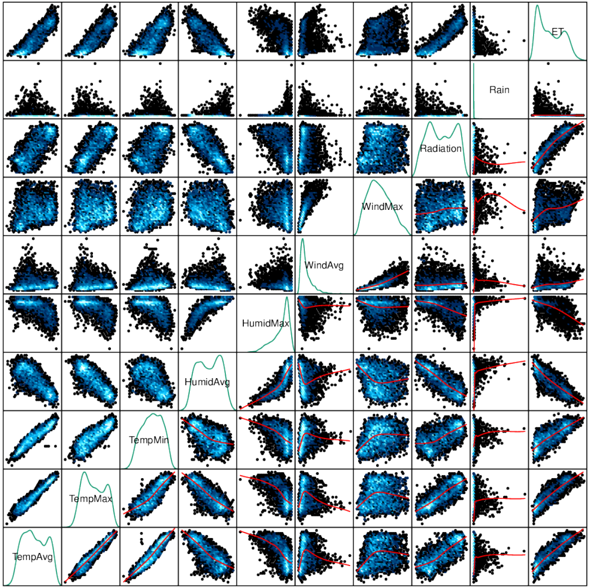
Gráficos boxplot



Gráficos de dispersión



Matrices de gráficos de dispersión



Desviación entre modelo y observación

- Sea O el conjunto de observaciones (medidas) de una variable aleatoria.

$$\mathbf{O} = \{o_1 \dots o_n\}$$

- Sea M el conjunto de resultados de un modelo que aproxima el comportamiento de la variable medida.

$$\mathbf{M} = \{m_1 \dots m_n\}$$

- La desviación entre modelo y observación es:

$$\mathbf{D} = \mathbf{M} - \mathbf{O} = \{(m_1 - o_1) \dots (m_n - o_n)\} = \{d_1 \dots d_n\}$$

Exactitud (*bias*) y Precisión (*variance*)

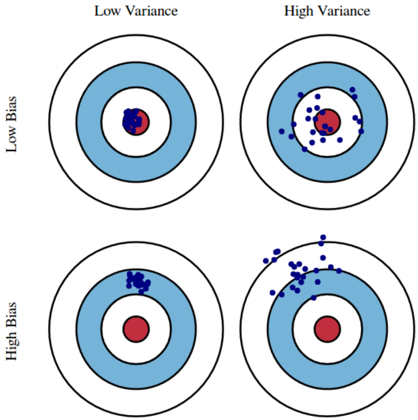


Fig. 1 Graphical illustration of bias and variance.

http://scott.fortmann-roe.com/docs/BiasVariance.html

Estimadores frecuentes: MBD y RMSD

- Mean Bias Difference (MBD), diferencia media (indica si el modelo sobreestima o subestima):

$$MBE = \overline{\mathbf{D}} = \overline{\mathbf{M}} - \overline{\mathbf{O}} = \frac{1}{n}\sum_{i=1}^{n}(m_i - o_i)$$

- Root Mean Square Difference (RMSD), diferencia cuadrático media:

$$RMSD = \left(\frac{1}{n}\sum_{i=1}^{n} d_i^2\right)^{1/2} = \left(\frac{1}{n}\sum_{i=1}^{n}(m_i - o_i)^2\right)^{1/2}$$

Estimadores frecuentes: MBE y RMSD

El RMSD agrega información del promedio y la varianza de la diferencia:

$$RMSD^2 = \overline{\mathbf{D}}^2 + \sigma_{\mathbf{D}}^2$$

donde la varianza de la diferencia (unbiased RMSD) se calcula:

$$\sigma_{\mathbf{D}}^2 = \frac{1}{n} \sum_{i=1}^{n} (d_i - \overline{\mathbf{D}})^2$$

Otros estimadores: MAD

- Mean Absolute Deviation (MAD):

$$MAD = \frac{1}{n}\sum_{i=1}^{n} |d_i| = \frac{1}{n}\sum_{i=1}^{n} |m_i - o_i|$$

- El RMSD no es robusto (un error puntual puede distorsionar el estimador) y depende del número de muestras:

$$MAD \leq RMSD \leq n^{1/2} MAD$$

Otros estimadores: t y d

- t de Student (valores pequeños indican buen comportamiento del modelo)
  - Permite añadir intervalos de confianza a las diferencias entre modelo y observación

$$t = \left(\frac{(n-1)MBD^2}{RMSD^2 - MBD^2} \right)^{1/2}$$

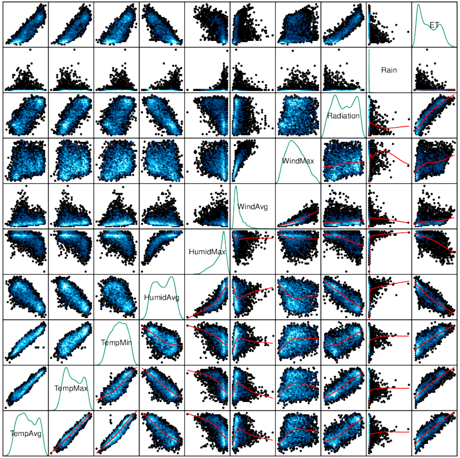
- d_1: Índice de concordancia de Willmott.
  - Limitado entre 0 (ausencia de concordancia) y 1 (concordancia total).
  - Robusto frente a *outliers*.

$$d_1 = 1 - \frac{\sum_{i=1}^{n} |m_i - o_i|}{\sum_{i=1}^{n} \left(\left| m_i - \overline{\mathbf{O}} \right| + \left| o_i - \overline{\mathbf{O}} \right| \right)}$$

Correlación

El coeficiente de correlación entre dos conjuntos de datos es una medida numérica de la relación **lineal** entre los dos conjuntos (si la relación no es lineal, este coeficiente no sirve):

$$r = \frac{1}{n-1} \cdot \sum_{i=1}^{n} \left(\frac{o_i - \overline{\mathbf{O}}}{\sigma_{\mathbf{O}}} \right) \cdot \left(\frac{m_i - \overline{\mathbf{M}}}{\sigma_{\mathbf{M}}} \right)$$



Diagramas de Taylor

- Desarrollando $\sigma_{\mathbf{D}}^2$ y teniendo en cuenta la definición de r:

$$\sigma_{\mathbf{D}}^2 = \sigma_{\mathbf{O}}^2 + \sigma_{\mathbf{M}}^2 - 2 \cdot \sigma_{\mathbf{O}} \cdot \sigma_{\mathbf{M}} \cdot r$$

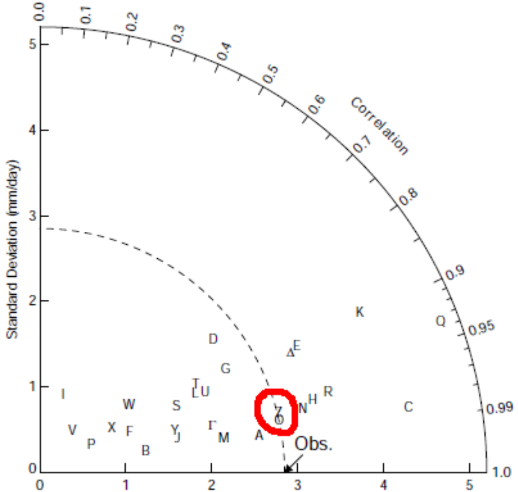
- Esta relación es semejante a la ley de los cosenos (c, a, b son lados de un triángulo y ϕ es el ángulo opuesto al lado c):

$$c^2 = a^2 + b^2 - 2 \cdot a \cdot b \cos \phi$$

Diagramas de Taylor

$$\sigma_{\mathbf{D}}^2 = \sigma_{\mathbf{O}}^2 + \sigma_{\mathbf{M}}^2 - 2 \cdot \sigma_{\mathbf{O}} \cdot \sigma_{\mathbf{M}} \cdot r$$

- $\sigma_{\mathbf{D}}^2$: Distancia al origen
- $\sigma_{\mathbf{O}}^2$: Eje horizontal
- $\sigma_{\mathbf{M}}^2$: Eje vertical
- r: acimut



Target Diagram

- Emplea la relación entre $RMSD$, $\sigma_{\mathbf{D}}^2$, y $\overline{\mathbf{D}}$, normalizadas con $\sigma_{\mathbf{O}}$:

$$RMSD' = RMSD/\sigma_{\mathbf{O}}$$

$$\sigma'_{\mathbf{D}} = \sigma_{\mathbf{D}}/\sigma_{\mathbf{O}}$$

$$\overline{\mathbf{D}}' = \overline{\mathbf{D}}/\sigma_{\mathbf{O}}$$

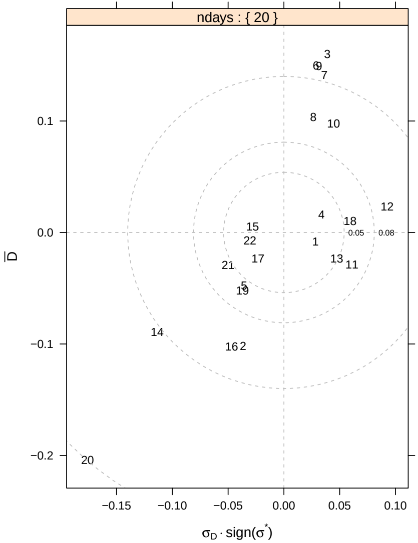
$$RMSD'^2 = \sigma'^2_{\mathbf{D}} + \overline{\mathbf{D}}'^2$$

- Incorporan el signo de la diferencia entre desviaciones estándar del modelo y las observaciones.

$$sign_{\sigma} = sign(\sigma_{\mathbf{M}} - \sigma_{\mathbf{O}})$$

Target Diagram

- $\sigma'_{\mathbf{D}}$ (con signo): Eje horizontal
- $\overline{\mathbf{D}}'$: Eje vertical
- $RMSD'^2$: Distancia al origen



Planteamiento

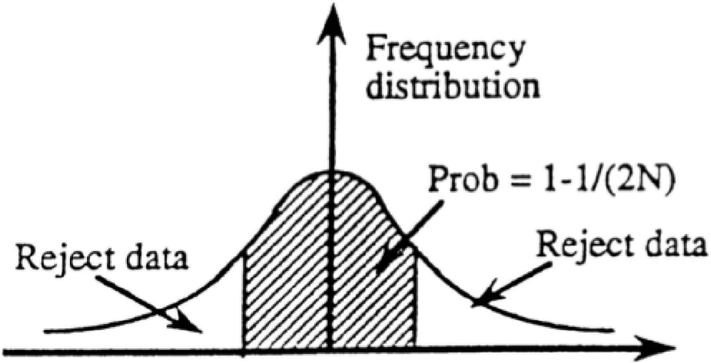
- En un sistema compuesto por diferentes **unidades idénticas**, el funcionamiento de todas ellas debe ser el mismo.
- Los **elementos reales** presentan **diferencias** (tolerancia de potencia, pérdidas de dispersión, suciedad), de forma que el comportamiento de cada unidad se desviará del comportamiento promedio.
- Si **no hay problemas de funcionamiento**, estas desviaciones no superarán un umbral determinado (**coherencia estadística**).
- Un análisis estadístico del funcionamiento del conjunto puede **identificar una unidad defectuosa** como aquella que se aparta significativamente del comportamiento promedio.

Coherencia estadística

- Una medida puede ser etiquetada como *outlier* si es poco probable que pertenezca a la misma distribución que el conjunto.
- En estadística hay métodos diversos para realizar este análisis:
  - Métodos gráficos
  - Teorema de Chebyshev.
  - Criterios de Pierce y Chauvenet
- Los criterios de Pierce y Chauvenet asumen **distribución gaussiana** (aceptable en nuestro contexto) y son **simples de implementar**, particularmente Chauvenet.

Método de Chauvenet

Una medida es un *outlier* si la probabilidad de obtener su desviación respecto de la media es inferior al inverso de 2 veces el número de elementos en el conjunto.



Método de Chauvenet

1. Supongamos un SFV compuesto por N unidades idénticas.
2. Sea $Y_{f,i}$ la medida de **productividad diaria** de una de esas unidades.
3. Se calcula la **media**, $\overline{Y}_f$, y la **desviación estándar**, σ_{Y_f}, del **conjunto**.
4. Se calcula la **distancia estadística** de cada unidad respecto del sistema total:

$$d_i = \frac{Y_{f,i} - \overline{Y}_f}{\sigma_{Y_f}}$$

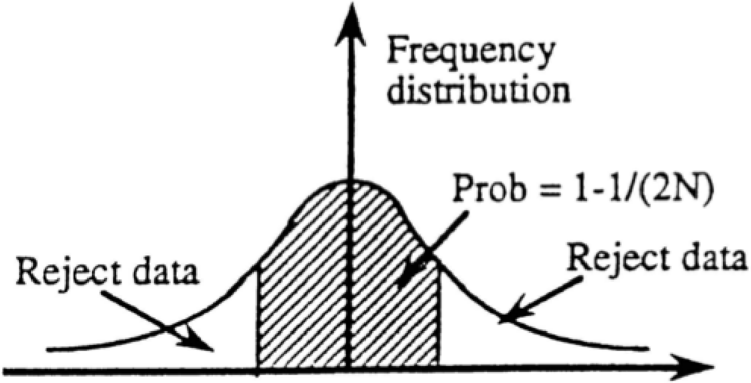
5. En una distribución gaussiana, se calcula la distancia estadística equivalente a la probabilidad límite de Chauvenet, $1/2N$, teniendo en cuenta que hay dos colas.
6. Aquellas observaciones que superan la distancia son **marcadas como outliers**.

Método de Chauvenet

- Por ejemplo, para un sistema de 10 unidades, la probabilidad es $1/20$.
- A cada cola le corresponde la mitad, $1/40 = 0.025$.
- Por tanto, el límite es $d_{max} = -1.9599$ (cola izquierda, valores por debajo del conjunto).

$$d_i = \frac{Y_{f,i} - \overline{Y}_f}{\sigma_{Y_f}}$$

$$|d_i| > |d_{max}|$$



Métodos gráficos

- Sea $Y_{f,i}(t)$ la serie temporal de la productividad diaria de **una unidad** del SFV.
- Sea $\widetilde{Y}_f(t)$ la serie temporal de la **mediana** de la productividad diaria del SFV **completo**.
- Tomando $\widetilde{Y}_f(t)$ como modelo o referencia, podemos emplear los diagramas de Taylor y Target para detectar unidades defectuosas durante períodos de tiempo.