

Técnicas de tratamiento estadístico

Un breve resumen de lo que se está haciendo en el DPAC y de lo que se podrá hacer con el archivo final de GAIA

L.M. Sarro^{1,2}

¹Departamento de Inteligencia Artificial - UNED

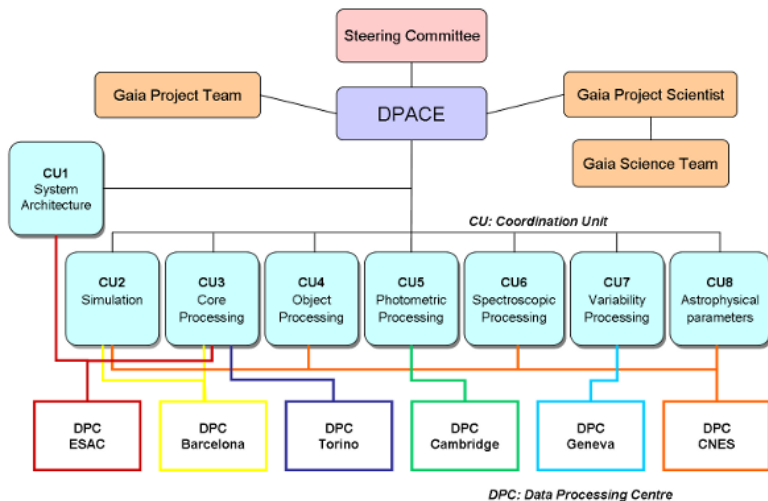
²Grupo de Minería de Datos del Observatorio Virtual Español

Otoño 2009 / Gaia: la Galàxia en un Petabyte - XXV
Trobades científiques de la Mediterrànea

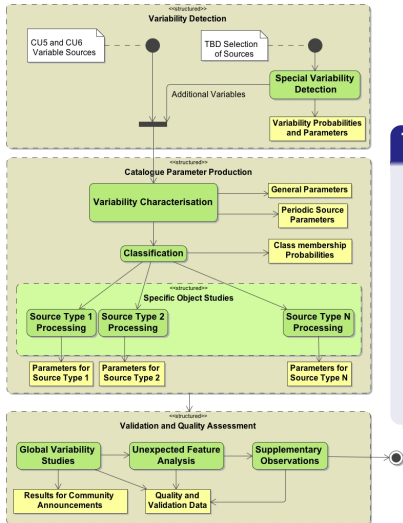
Outline

- 1 Estadística en el DPAC
 - DPAC y la construcción del archivo final de la misión
 - ¿Y después?

Estadística en el DPAC



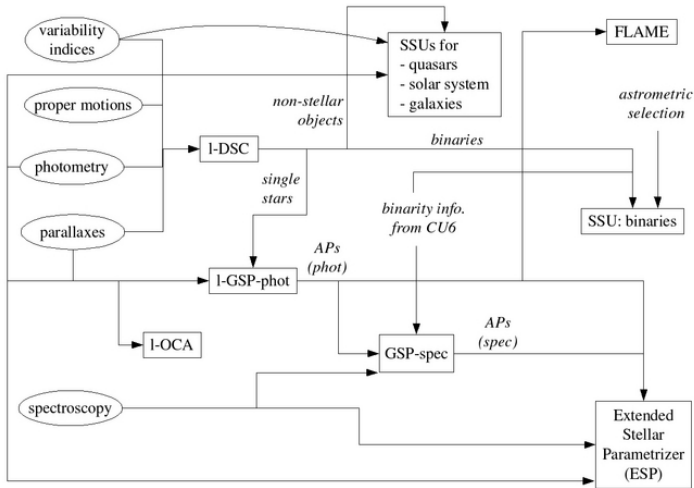
CU7



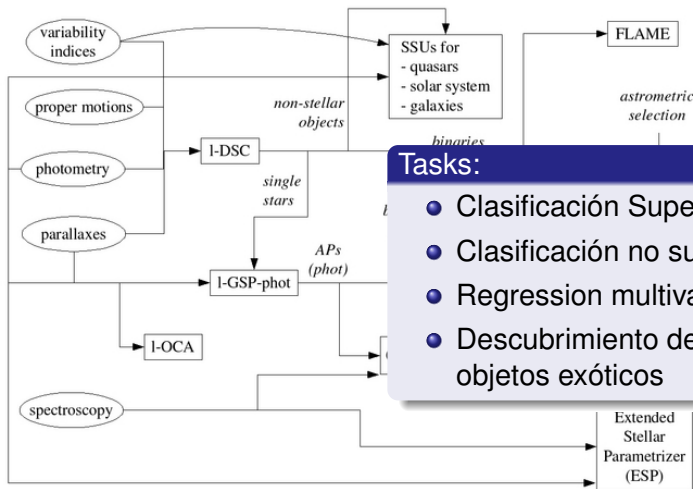
Tasks:

- Clasificación supervisada
- Clasificación no supervisada
- Descubrimiento de clases y objetos exóticos
- Análisis de sesgos
- Comparación de surveys

CU8



CU8



Tasks:

- Clasificación Supervisada
- Clasificación no supervisada
- Regression multivariante
- Descubrimiento de clases y objetos exóticos

Técnicas modernas de estadística multivariante.

Poco nuevo bajo el Sol...

- Extracción de características: wavelets, PCA, ICA, NMF
- Clasificación supervisada y regresión multivariante
- Estimación Paramétrica Bayesiana
- Técnicas de agrupamiento o *clustering* (Clasificación no supervisada)
- Técnicas de detección de objetos exóticos
- Control de calidad: comparación de surveys.

Técnicas modernas de estadística multivariante.

Clasificación supervisada o regresión

$$R^n \rightarrow g \rightarrow R^m \quad (1)$$

donde g es un modelo de aprendizaje estadístico a partir de un conjunto de entrenamiento:

- Redes neuronales artificiales
- Máquinas de Vectores Soporte
- Redes Bayesianas
- Regresión logística, ridge...
- Árboles de decisión
- Y más...

Más detalle...

Elements of statistical learning, Hastie & Tibshirani, Springer
Data Mining, Christopher M. Bishop, Oxford.

Técnicas modernas de estadística multivariante.

Estimación bayesiana de parámetros

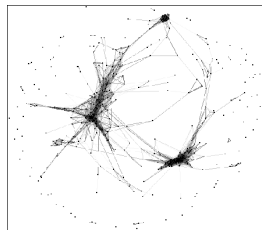
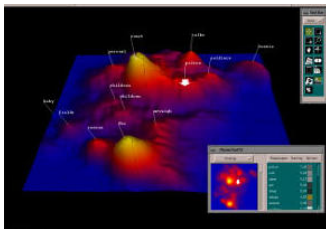
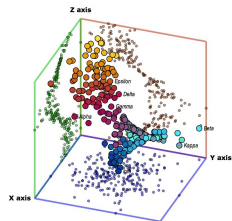
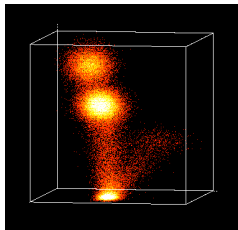
$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})} \quad (2)$$

Ventajas:

- Cálculo riguroso y bien fundamentado de los errores en la estimación (matriz de Fisher)
- Proporciona distribuciones de probabilidad para los parámetros, no valores únicos.
- Selección de modelos, factor de Bayes, MCMC, nested sampling...

Ver [*Statistical techniques in cosmology*](#), Heavens, astro-ph
Junio 2009

Técnicas de agrupamiento o *clustering*



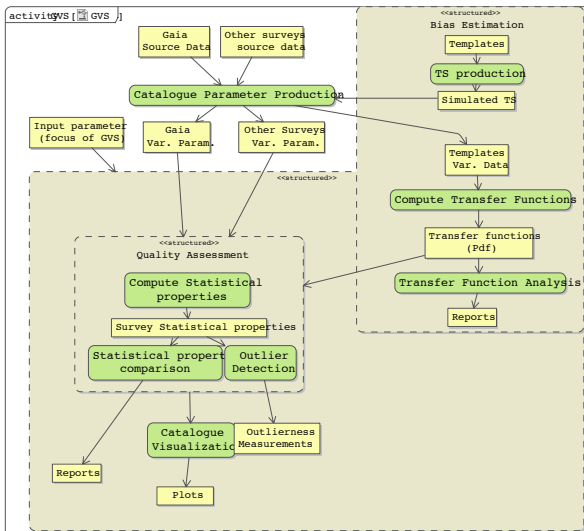
Técnicas de agrupamiento o *clustering*

¿Por qué hacemos clustering?

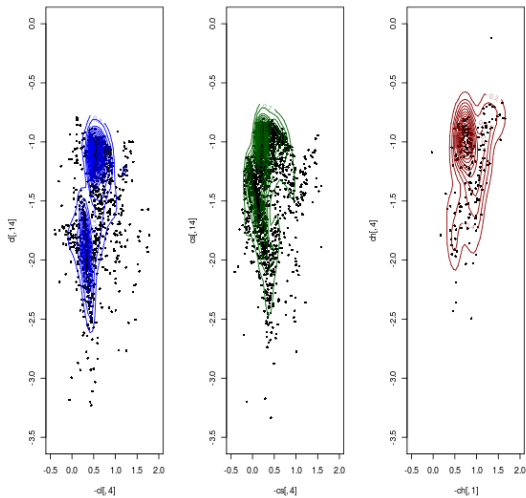
Porque gran parte del trabajo de clasificación/regresión de CU7 y CU8 se basa en la hipótesis (con gran probabilidad falsa) de que conocemos todas las clases de objetos.

Si durante la misión descubrimos una nueva clase de objetos, habrá que rediseñar y rehacer las cadenas de procesado.

Control de calidad: comparación de surveys



Control de calidad: comparación de surveys



Desafíos de GAIA: cómo explotar adecuadamente el archivo final.

¿Algo nuevo bajo el Sol?

¿Y después?

Es importante recordar que el DPAC **no hace ciencia**.

Por lo tanto, es importante dominar **o desarrollar** las técnicas (estadísticas y de todo tipo) óptimas para la explotación científica del archivo.