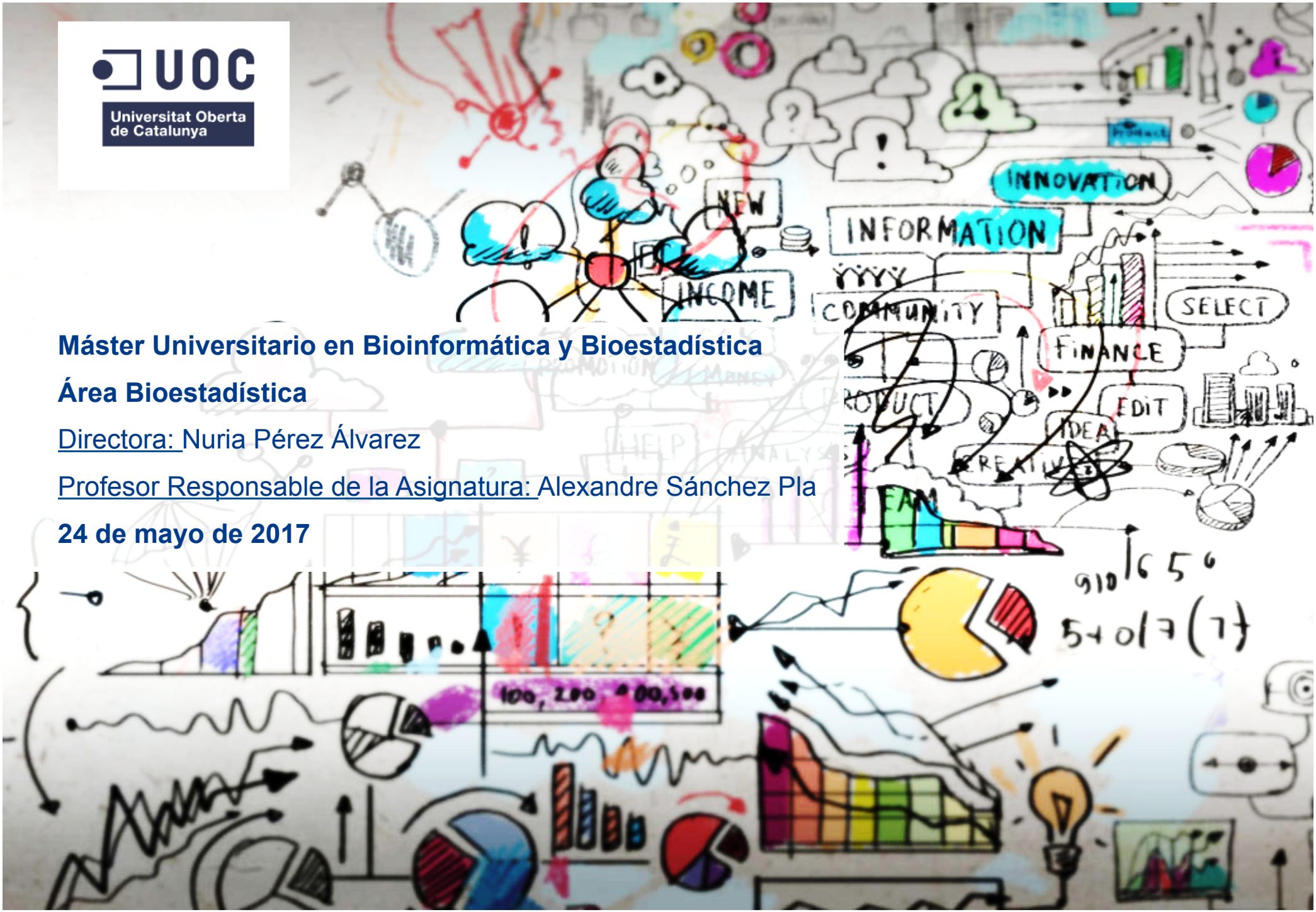
Trabajo de Fin de Máster

Estrategias para el tratamiento de datos faltantes ("missing data") en estudios con datos longitudinales

Román Octavio Calafati



Máster Universitario en Bioinformática y Bioestadística

Área Bioestadística

Directora: Nuria Pérez Álvarez

Profesor Responsable de la Asignatura: Alexandre Sánchez Pla

24 de mayo de 2017

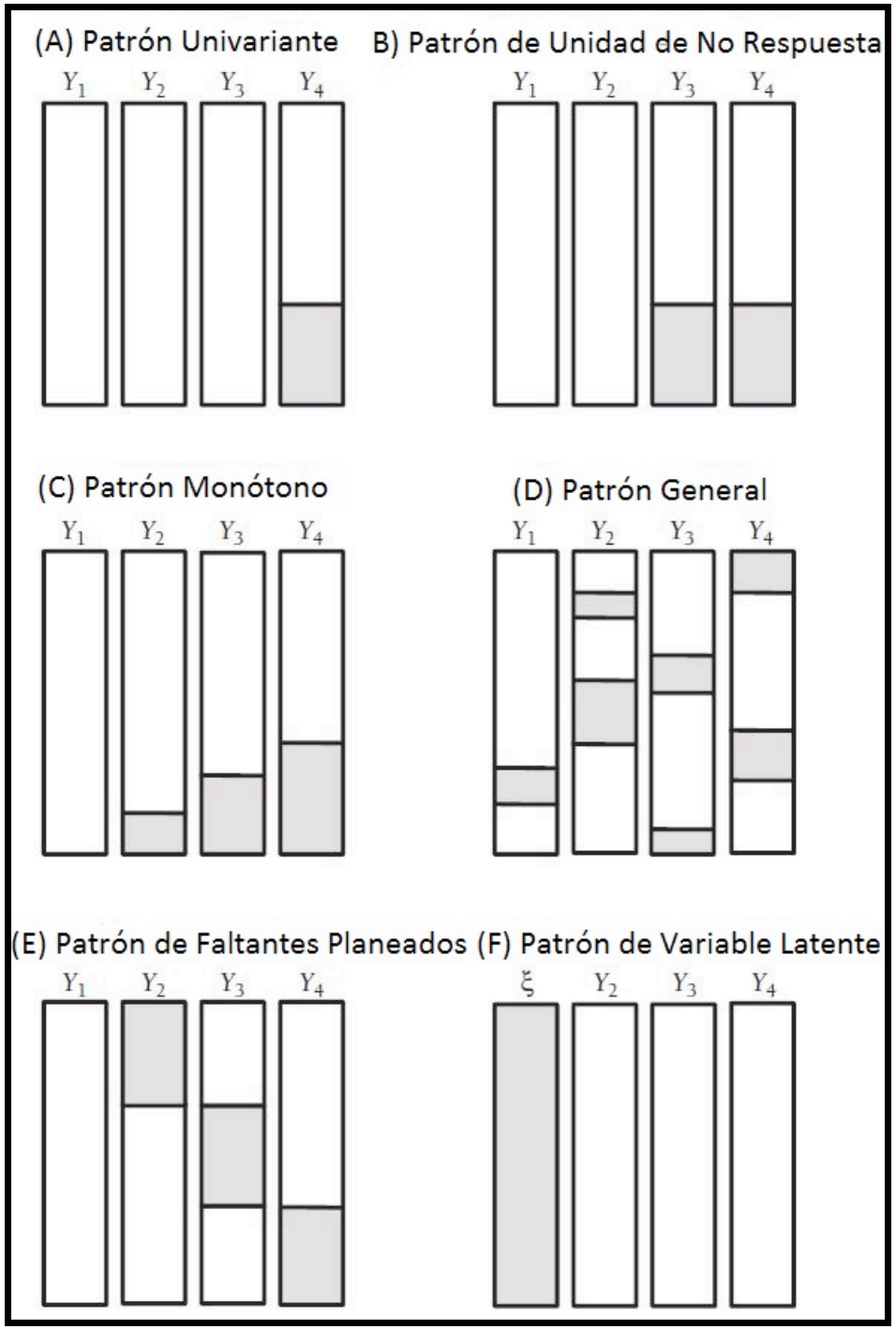
Introducción a la casuística de datos faltantes

Contexto	Evento
Actualmente existen muchos estudios en los cuales se realizan análisis estadísticos, y muchos de ellos contienen datos faltantes.	▪ Los investigadores utilizan métodos "ad hoc" para tratar los datos faltantes.
Los motivos por los cuales se producen datos faltantes son variados.	▪ Dado el precio de la recolección de los datos, la mayoría de las veces los estudios no pueden ser reiniciados.
Se hace necesario "tratar" los datos faltantes.	▪ Existen varios métodos disponibles para el tratamiento de datos faltantes.

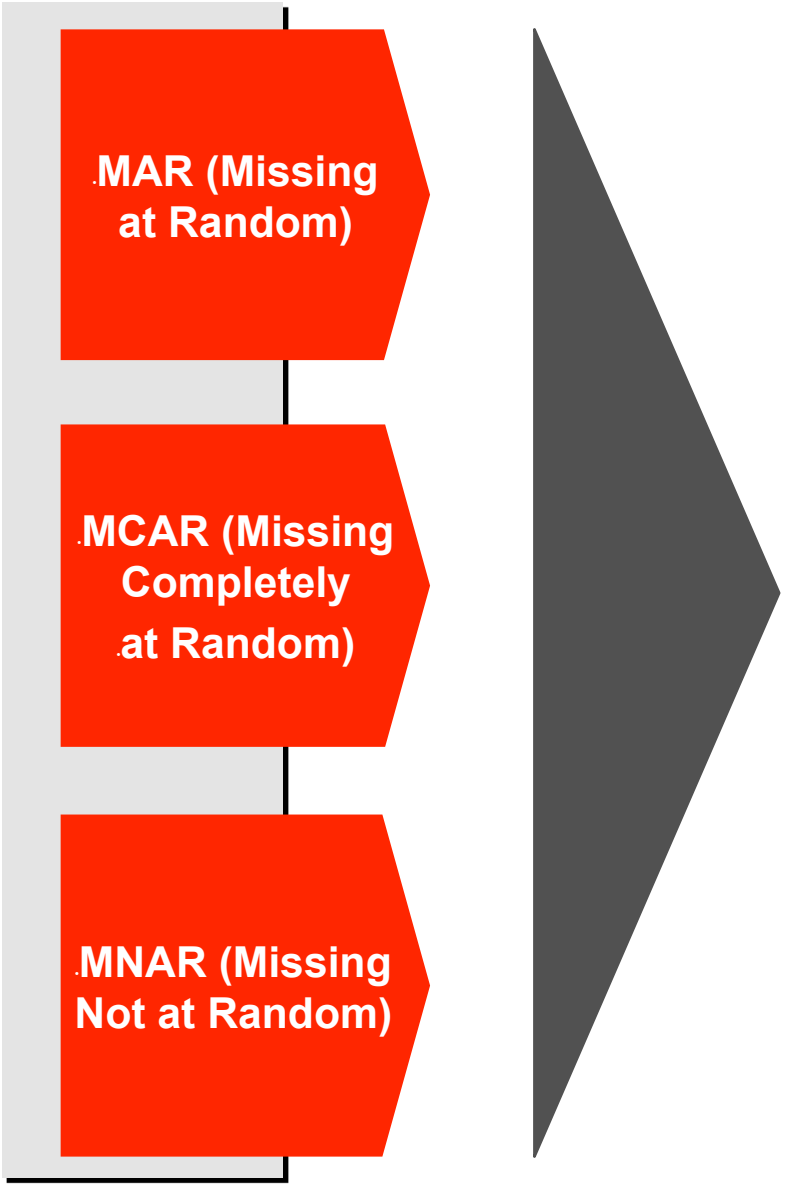
Ejemplo de una tabla con datos faltantes

Datos sobre selección de empleados		
IQ	Bienestar psicológico	Performance en el trabajo
78	13	-
84	9	-
84	10	-
85	10	-
87	-	-
91	3	-
92	12	-
94	3	-
94	13	-
96	-	-
99	6	7
105	12	10
105	14	11
106	10	15
108	-	10
112	10	10
113	14	12
115	14	14
118	12	16
134	11	12

Patrones típicos de datos faltantes



Sistema de clasificación actual para los datos faltantes, de acuerdo a Rubin et al. (1976)



Calificaciones de performance en el trabajo con valores faltantes MCAR, MAR y MNAR				
	Calificaciones de performance en el trabajo			
IQ	Completo	MCAR	MAR	MNAR
78	9	-	-	9
84	13	13	-	13
84	10	-	-	10
85	8	8	-	-
87	7	7	-	-
91	7	7	7	-
92	9	9	9	9
94	9	9	9	9
94	11	11	11	11
96	7	-	7	-
99	7	7	7	-
105	10	10	10	10
105	11	11	11	11
106	15	15	15	15
108	10	10	10	10
112	10	-	10	10
113	12	12	12	12
115	14	14	14	14
118	16	16	16	16
134	12	-	12	12

Métodos de eliminación de datos faltantes

Eliminación Pairwise (Análisis de caso disponible)

Al realizar algún cálculo estadístico, **elimina sólo la pareja asociada al valor faltante**, por ejemplo, en el cálculo de la correlación toma en cuenta los valores en azul:

Pairwise Deletion		
X	pulse	resp_rate
1	66	28
2	88	20
3	40	24
4	164	84
5	104	35
6		
7	48	16
8	60	
9	80	36
10	90	
-0.057790404		

Eliminación Listwise (caso completo)

Al realizar algún cálculo estadístico, **elimina la observación completa**, si la misma contiene algún dato faltante, por ejemplo, en el cálculo de la correlación toma en cuenta los valores en azul:

Listwise Deletion (Complete Cases)		
X	pulse	resp_rate
1	66	28
2	88	20
3	40	24
4	164	84
5	104	35
6		
7	48	16
8	60	
9	80	36
10	90	
-0.02678537		

Métodos de imputación de datos faltantes

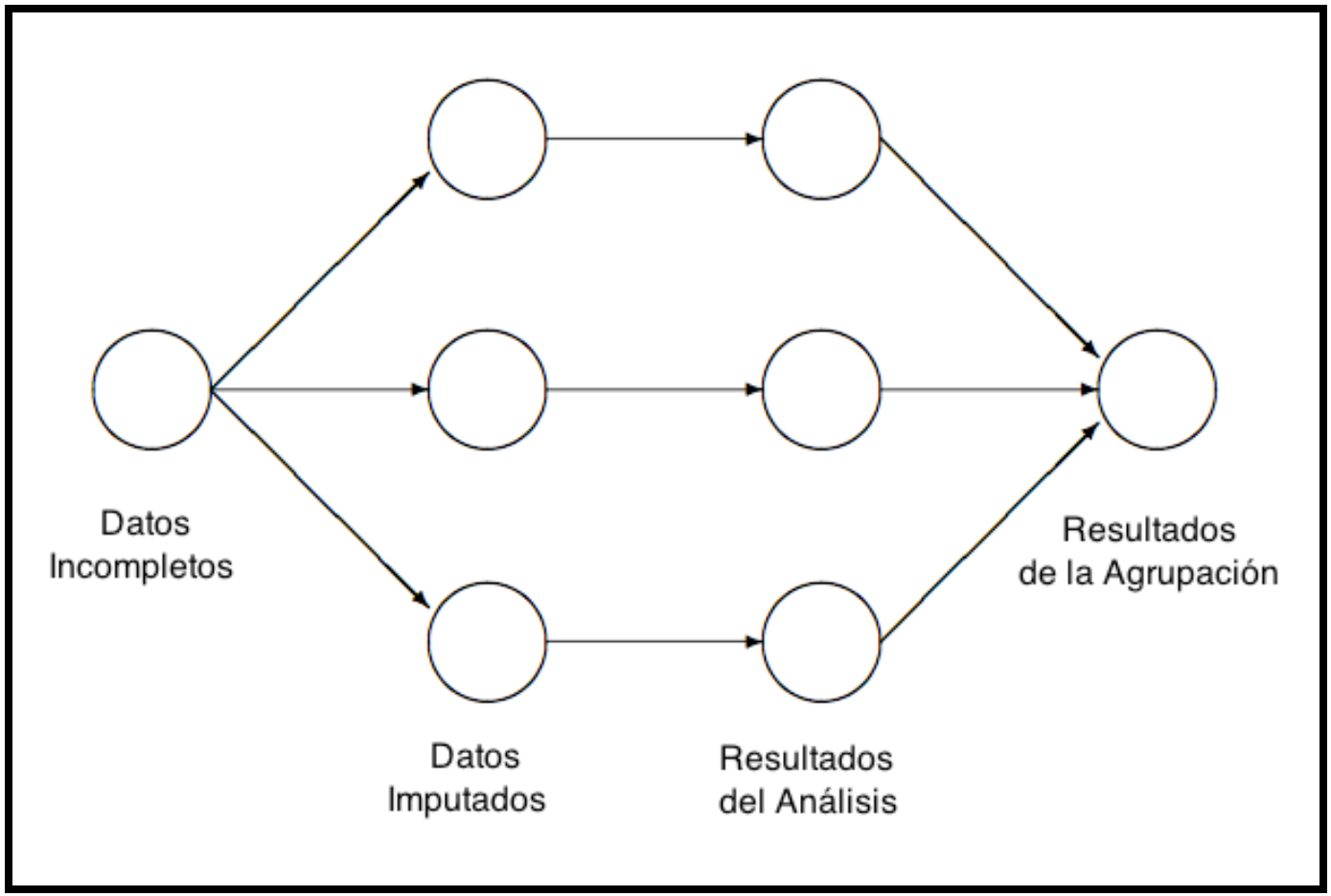
Imputación simple

- Hot-deck
- Cold-deck
- Media
- Moda
- Regresión
- Regresión estocástica
- LOCF

Imputación múltiple

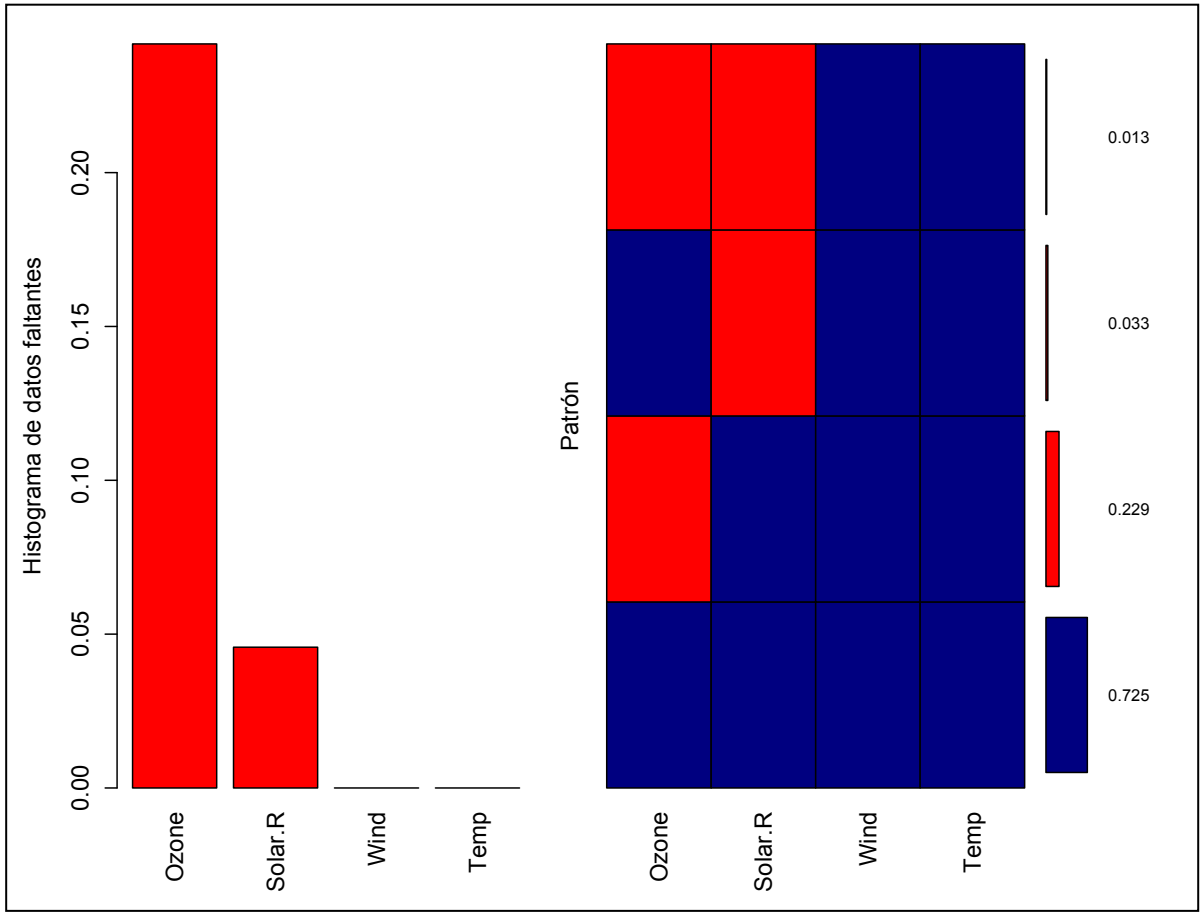
- Se utiliza un modelo de imputación, que reintroduce la variabilidad original de los datos imputados

Pasos principales en la imputación múltiple



Ejemplo utilizando el dataset "Airquality"

	Ozone	Solar.R	Wind	Temp	Month	Day
1	41	190	7.4	67	5	1
2	36	118	8.0	72	5	2
3	12	149	12.6	74	5	3
4	18	313	11.5	62	5	4
5	NA	NA	14.3	56	5	5
6	28	NA	14.9	66	5	6
7	23	299	8.6	65	5	7
8	19	99	13.8	59	5	8
9	8	19	20.1	61	5	9
10	NA	194	8.6	69	5	10
11	7	NA	6.9	74	5	11
12	16	256	9.7	69	5	12
13	11	290	9.2	66	5	13



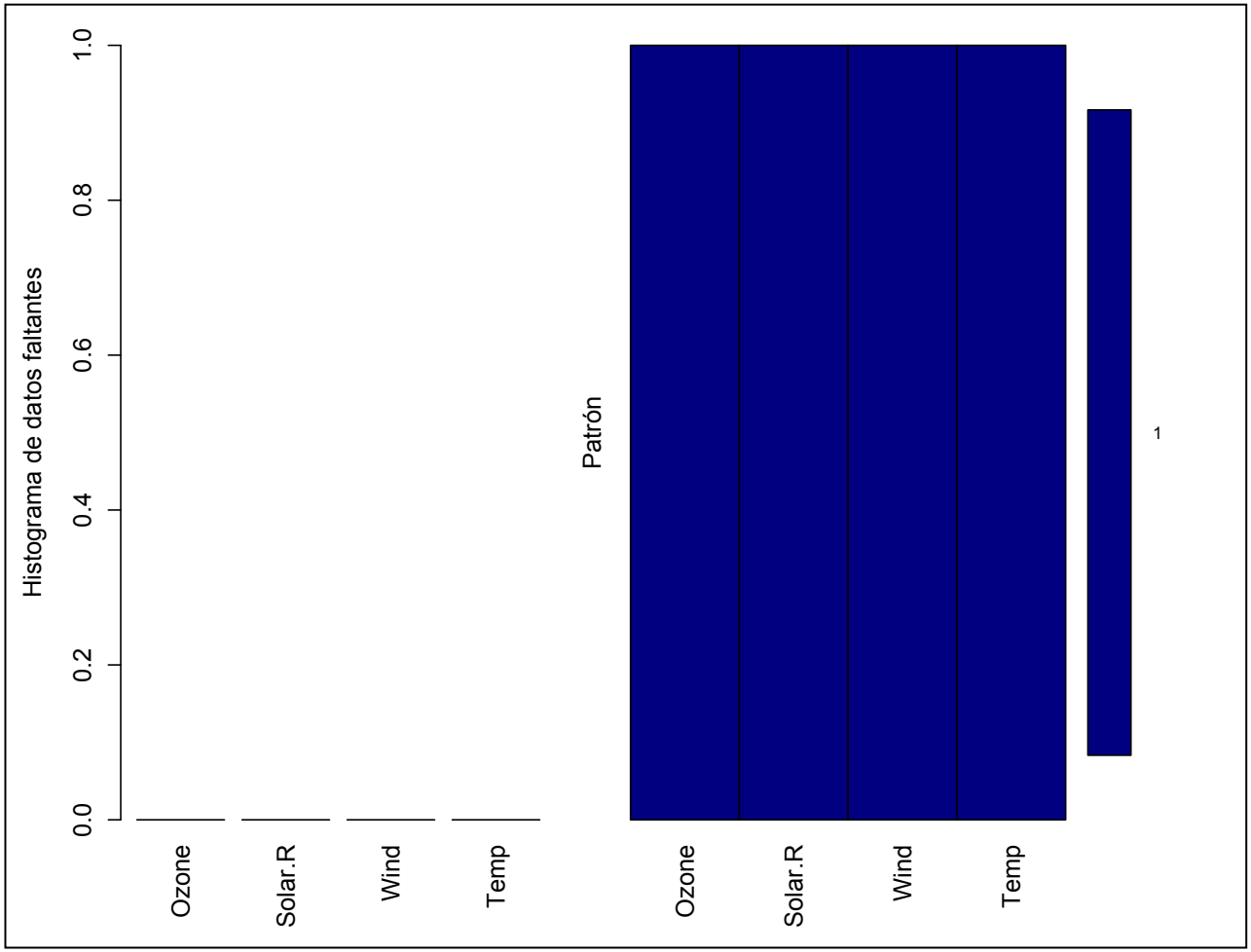
Datos importantes sobre el dataset Airquality (con 4 variables)

- 153 observaciones en total
- 111 observaciones completas y 42 incompletas

Resultados de la imputación realizada al dataset "Airquality"

	Ozone	Solar.R	Wind	Temp
1	41	190	7.4	67
2	36	118	8.0	72
3	12	149	12.6	74
4	18	313	11.5	62
5	NA	NA	14.3	56
6	28	NA	14.9	66
7	23	299	8.6	65
8	19	99	13.8	59
9	8	19	20.1	61
10	NA	194	8.6	69

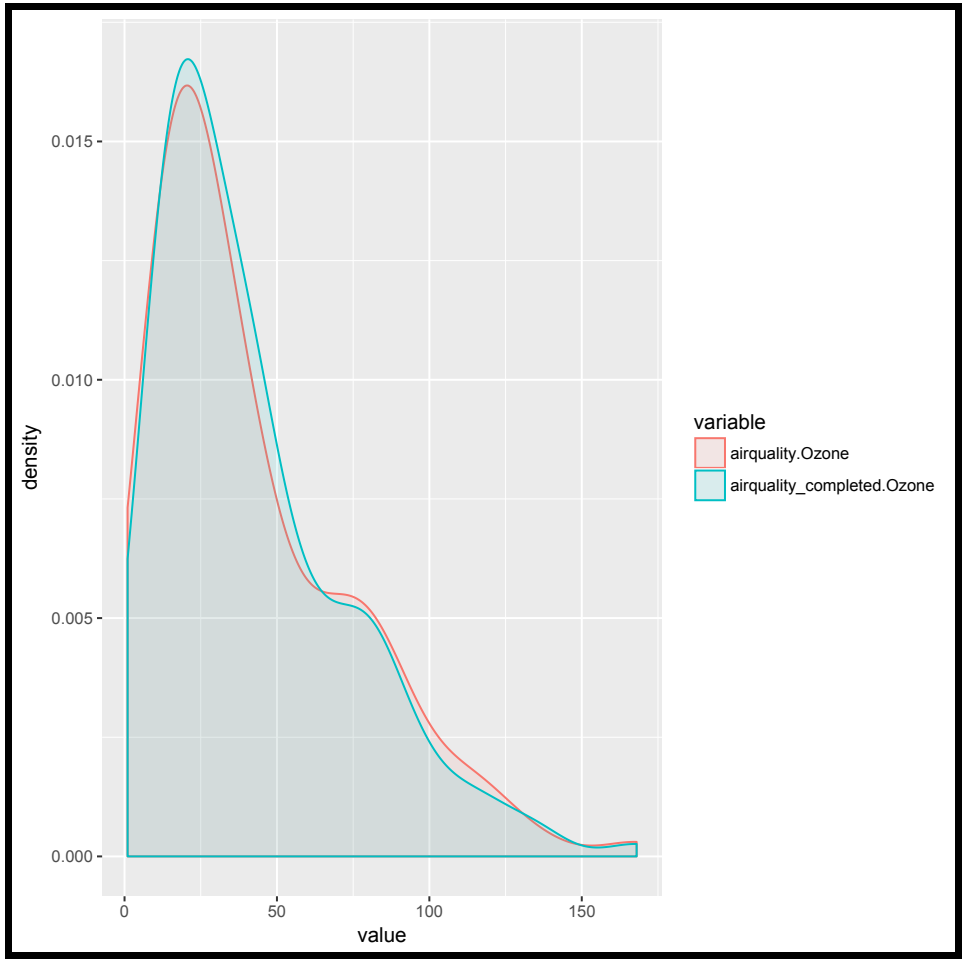
	Ozone	Solar.R	Wind	Temp
1	41	190	7.4	67
2	36	118	8.0	72
3	12	149	12.6	74
4	18	313	11.5	62
5	14	24	14.3	56
6	28	258	14.9	66
7	23	299	8.6	65
8	19	99	13.8	59
9	8	19	20.1	61
10	41	194	8.6	69

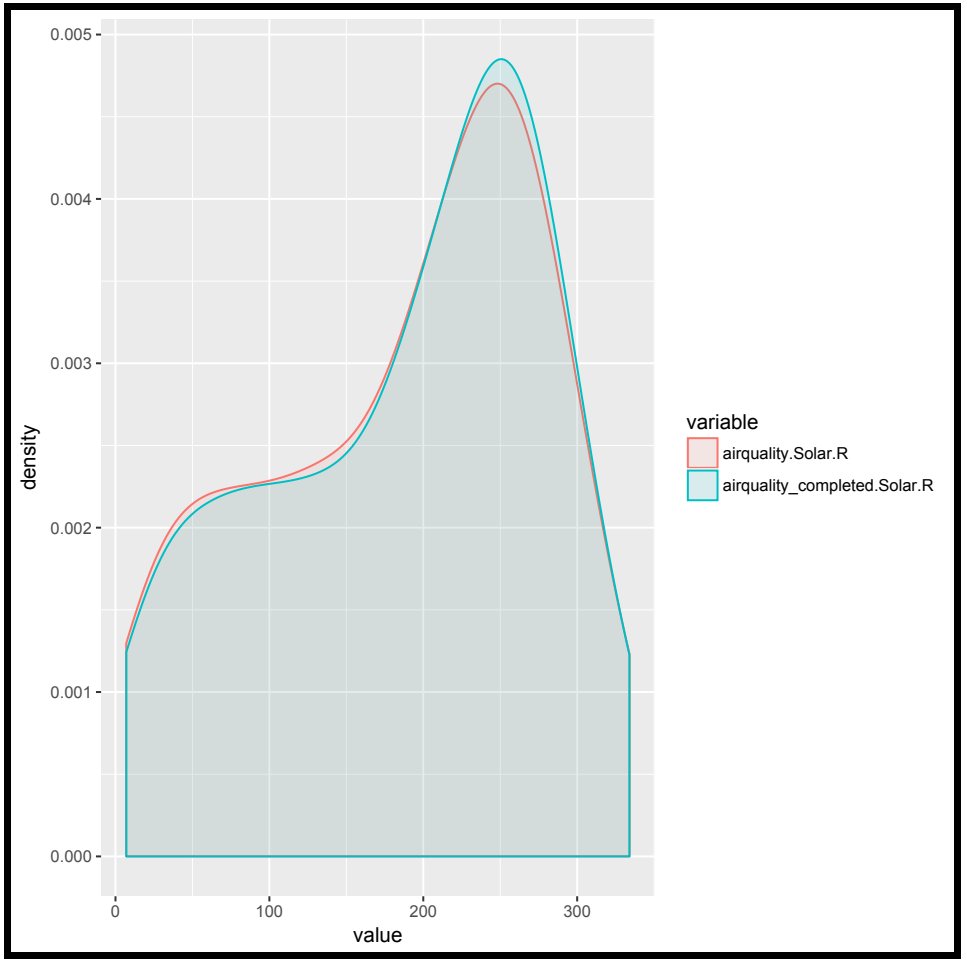


Datos importantes sobre el dataset Airquality (ya imputado)

- 153 observaciones en total
- 153 observaciones completas y 0 incompletas

Discusión y Conclusiones sobre el método PMM





Diversos estudios determinan que:

- El método PMM opera mejor que otros métodos cuando las muestras son grandes y no pequeñas (calcula con mayor precisión la variabilidad).
- Cuando la cantidad de datos faltantes se encuentra entre 10 % y 50 %, el método PMM obtiene mejores resultados que otros métodos. Si la cantidad de datos faltantes es menor al 5 %, diversos métodos obtienen los mismos resultados.

Ficheros entregados en este Trabajo de Fin de Máster

Objetivo: reproducibilidad

Descripción de los ficheros entregados

- Memoria del Trabajo
  - Ficheros:
    - "TFM_Memoria_Final_Alumno_Roman_Octavio_Calafati.doc"
    - "TFM_Memoria_Final_Alumno_Roman_Octavio_Calafati.pdf"
- Anexo de Prácticas
  - Ficheros:
    - "TFM_Anexo_de_Prácticas.rmd"
    - "TFM_Anexo_de_Prácticas.doc"
    - "TFM_Anexo_de_Prácticas.pdf"
- Anexo de Planificación
  - Ficheros:
    - "TFM_Anexo_de_Planificación.doc"
    - "TFM_Anexo_de_Planificación.pdf"
- Datasets utilizados
  - Ficheros:
    - "HOC.rds" (Dataset "Horse-Colic")
    - "AIR.rds" (Dataset "Air Quality"
    - "BOYS.rds" (Dataset "Boys")
    - "WALK.rds" (Dataset "Walking")
    - "NHAN.rds" (Dataset "Nhanes")
    - "PATT.rds" (Dataset "Pattern4")
- Presentación en Powerpoint
  - Ficheros:
    - "Presentación.ppt"
    - "Presentación.pdf"
    - "Presentación.Rmd" (Código fuente de los ejemplos desarrollados en la presentación)
- Vídeo con voz en off explicando la presentación en Powerpoint
  - Ficheros:
    - "Presentación en vídeo.mp4"