Instituto Meteorológico Nacional

Departamento de Redes Meteorológicas y Procesamiento de Datos

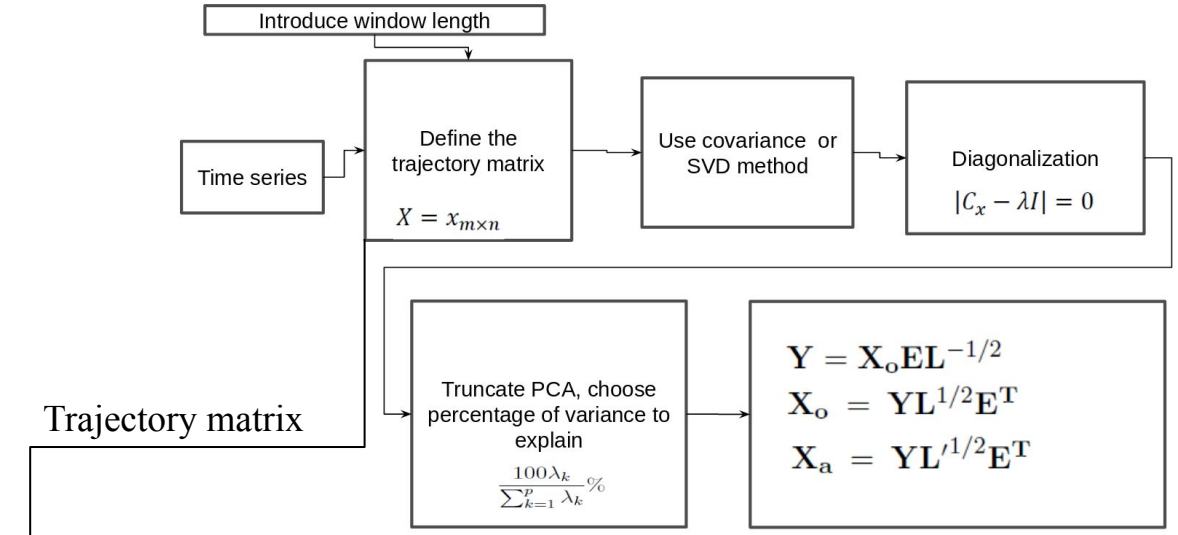
Relleno de datos horarios de temperatura superficial del aire usando ePCA

José Luis Araya

Objetivos

- Sistematizar una metodología de relleno para datos en resolución horaria.
- Proveer a diversos usuarios de datos meteorológicos rellenados y con reportes de datos mejorados.
- Encontrar las limitaciones de la red en términos de distribución de estaciones para la aplicación de métodos como el propuesto aquí.

¿Qué es ePCA?



Trajectory matrix

$$X = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix}$$

Extended cross-correlation matrix

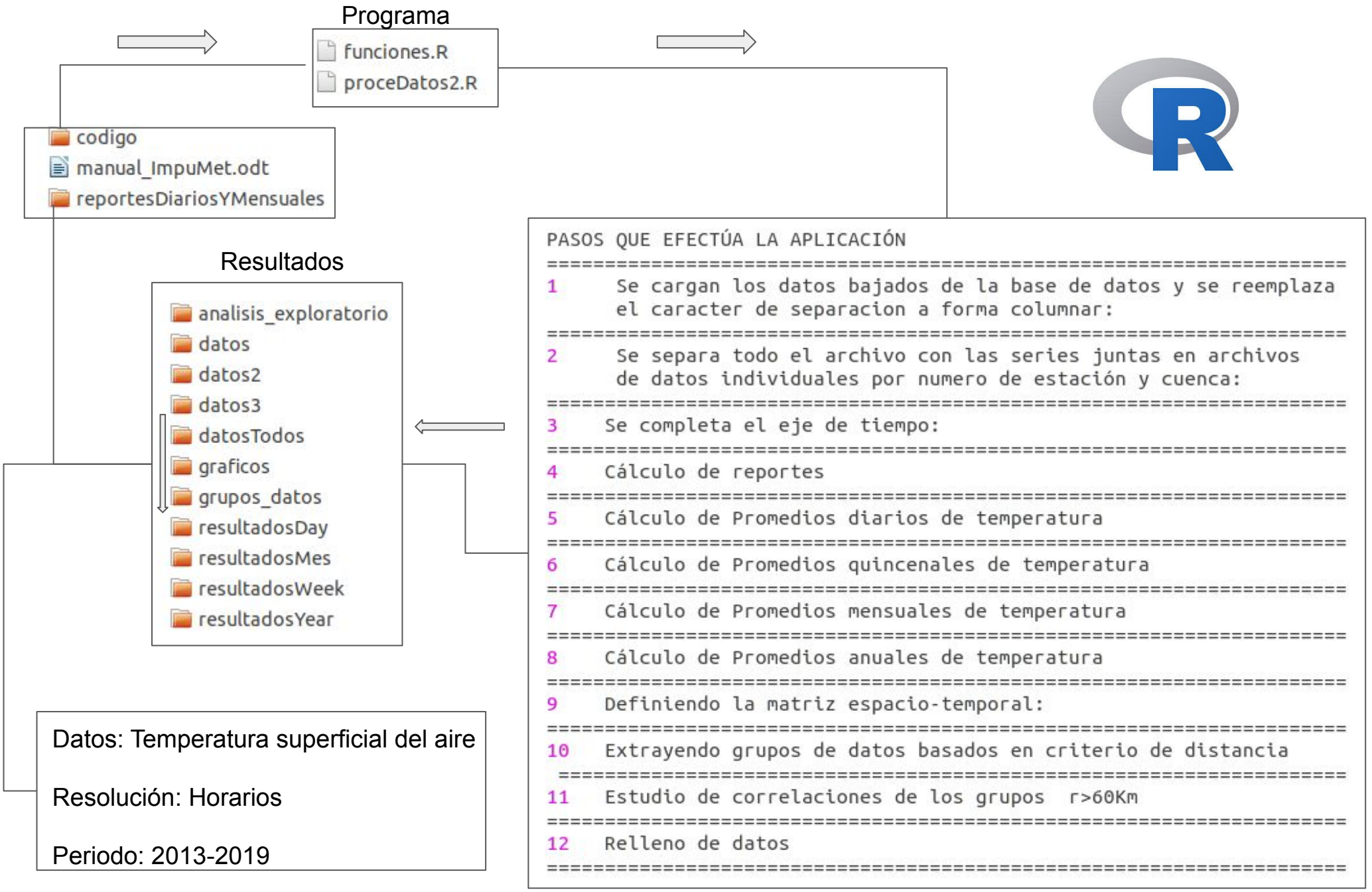
$$C = \frac{1}{n-M+1} XX^T = \begin{bmatrix} C_{1,1} & \cdots & C_{1,M} \\ \vdots & \ddots & \vdots \\ C_{M,1} & \cdots & C_{M,M} \end{bmatrix}$$

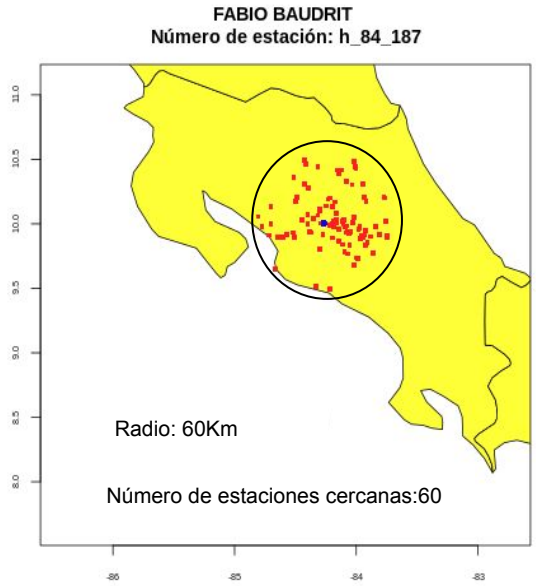
Criterios de selección de grupos

Un grupo aquí se define como un conjunto de series temporales con una estación de referencia, en torno a la cual se asocia otras estaciones usando diversos criterios. Los criterios de selección usados aquí son:

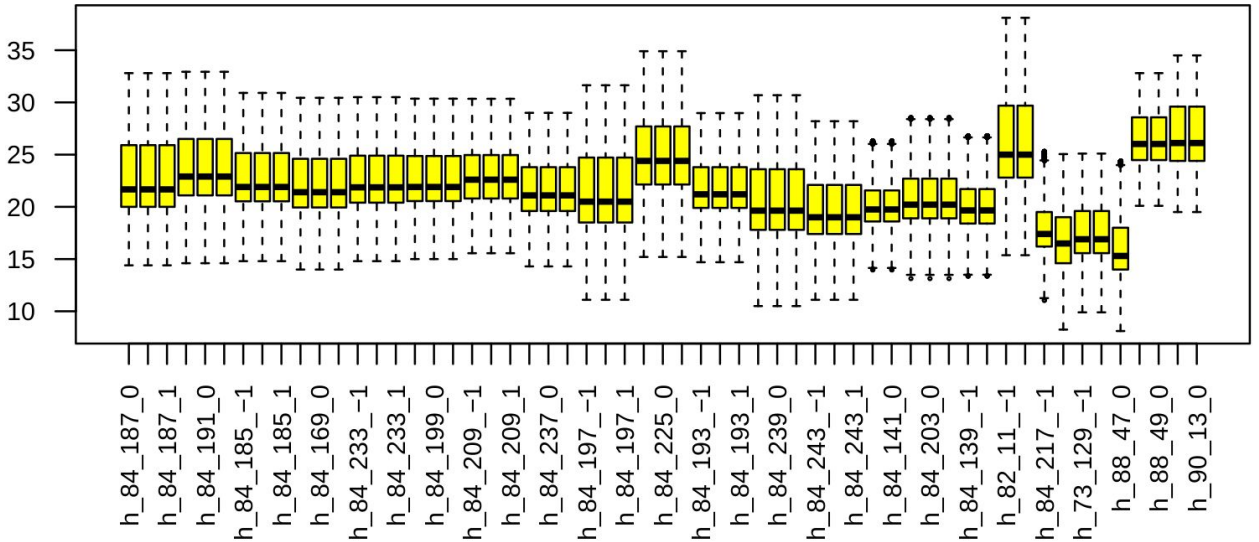
- Definir una estación de referencia
- Seleccionar estaciones existentes en 60 km a la redonda.
- Definir las estaciones circunvecinas con mayor correlación con respecto a la estación de referencia.

Aplicación que hace posible el relleno de datos de temperatura superficial horaria





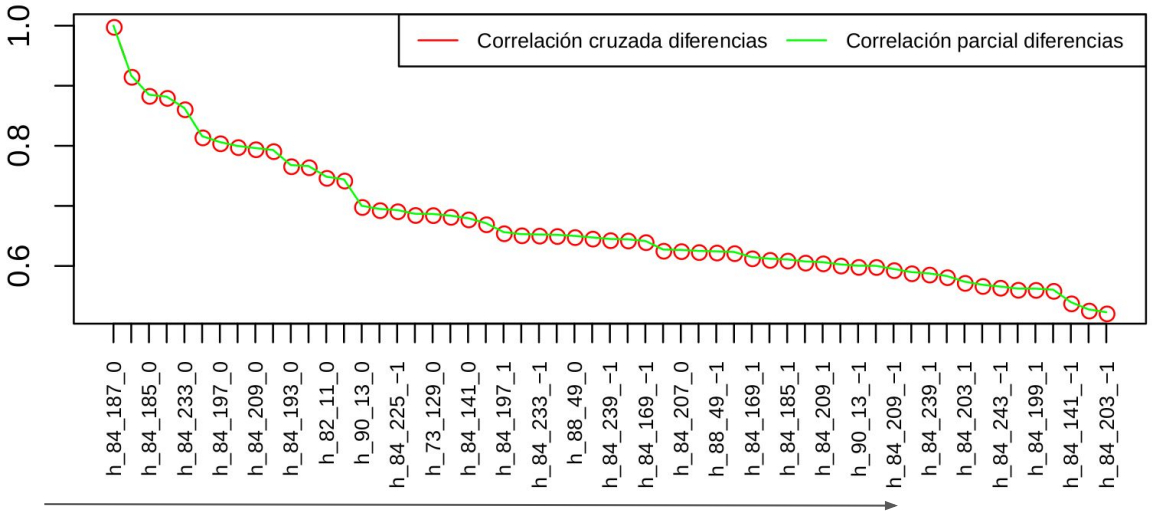
Ejemplo de grupos definidos para toda la nómina de estaciones automáticas



Cantagallo

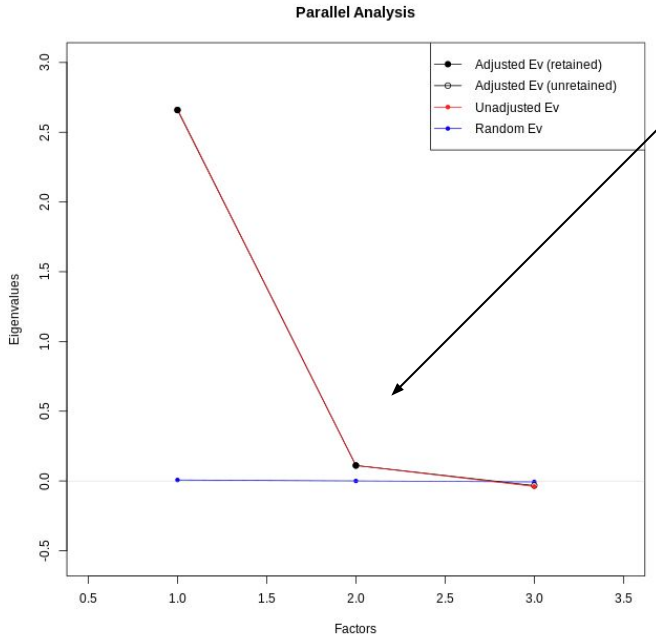
Código Distancia [km]

"distance_km"	
"h_71_15"	0
"h_71_23"	19.2531597941054
"h_71_21"	28.6792399133892
"h_69_675"	31.66741686128
"h_73_126"	32.8333719824343
"h_73_145"	33.1314277363696
"h_73_163"	33.1314277363696
"h_73_147"	33.4660740588714
"h_73_135"	34.5358147362531
"h_69_715"	36.1721589026049
"h_69_725"	36.625182644994
"h_69_697"	36.7005136653942
"h_69_681"	37.4901609778099
"h_73_159"	39.2821398199944
"h_69_721"	43.2119985024937
"h_69_731"	44.57004107393
"h_69_705"	44.8724880810187
"h_69_719"	48.0958103910716
"h_69_729"	50.1489729015823
"h_69_717"	52.360747585374
"h_69_709"	53.7057104808206
"h_73_143"	53.9097449726642
"h_73_167"	58.718496577243



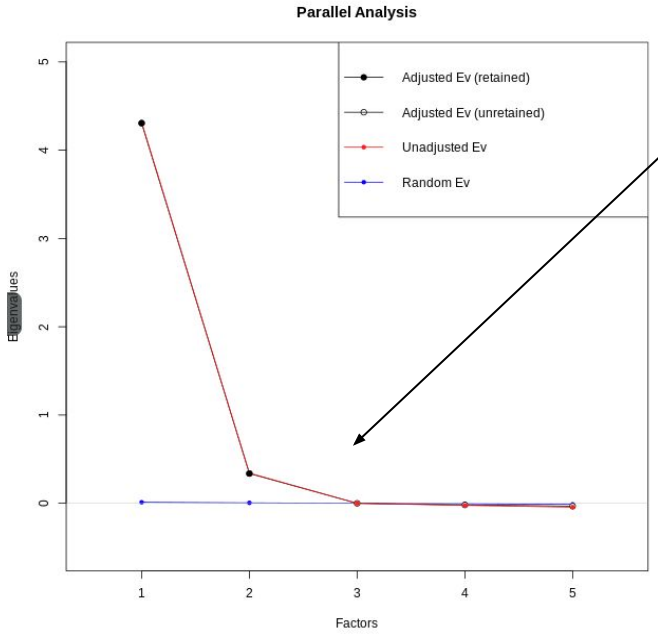
Estaciones cercanas ordenadas por distancia

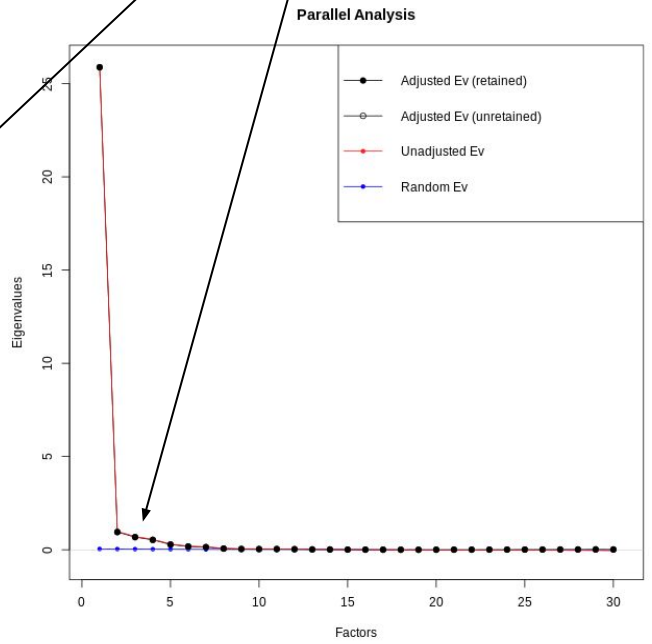
Definiendo el número óptimo de componentes principales

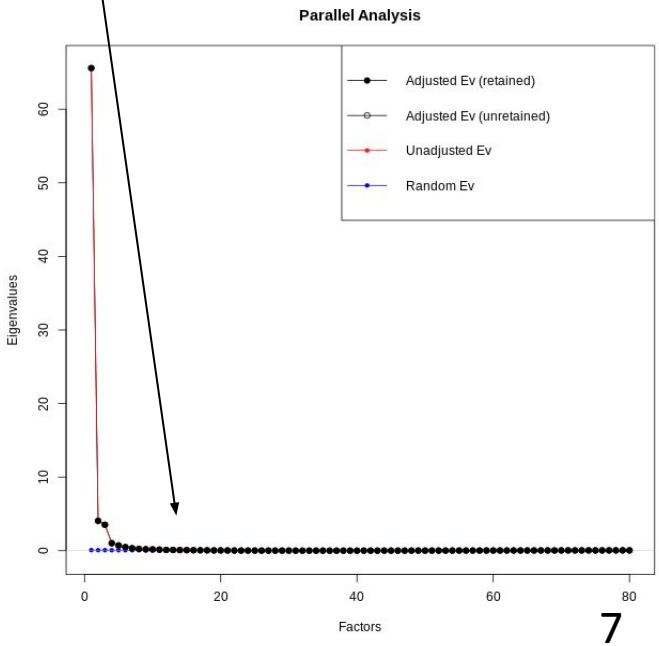


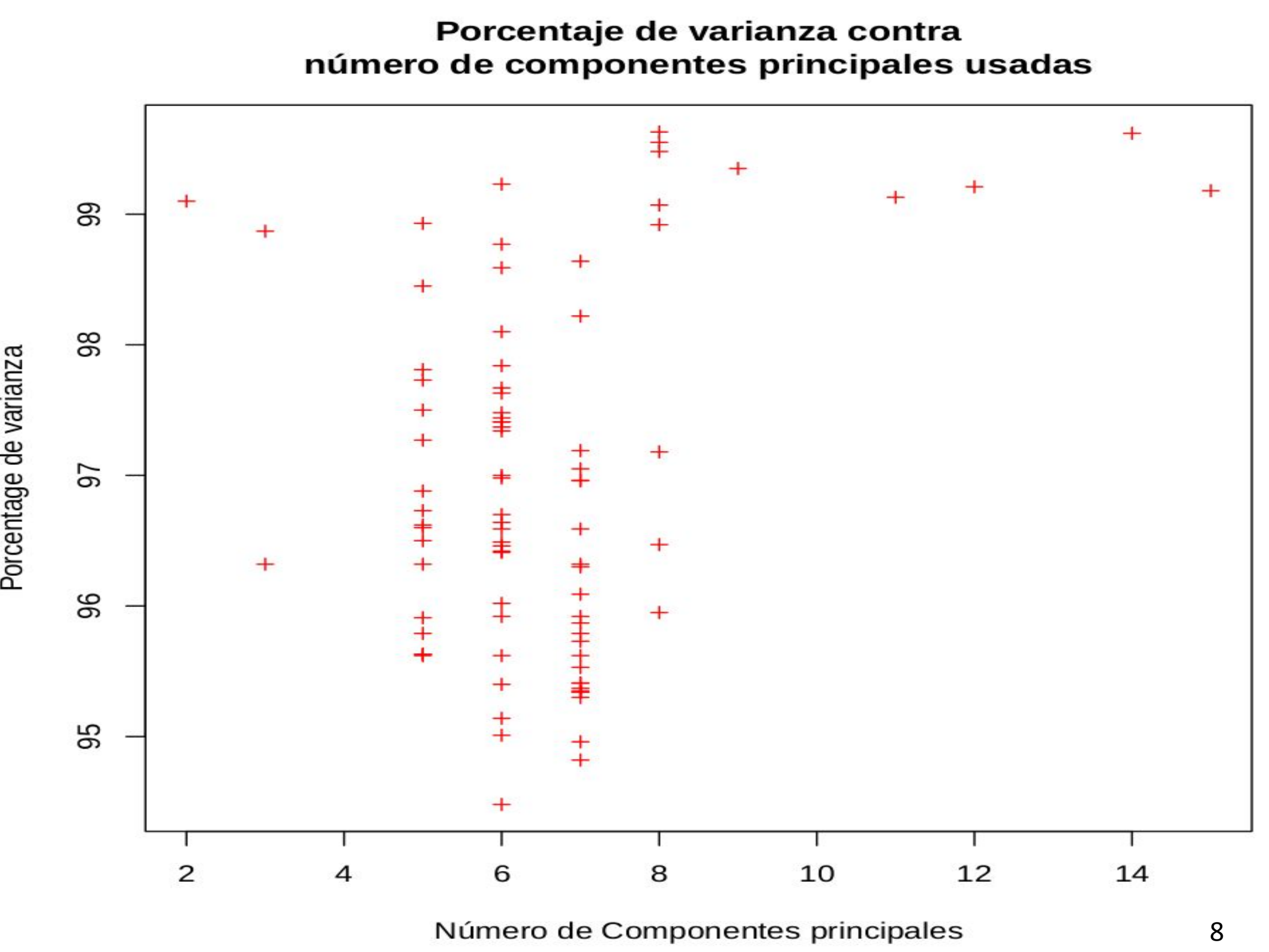
Con muy pocos factores la decisión del número de Componentes principales a retener es trivial.

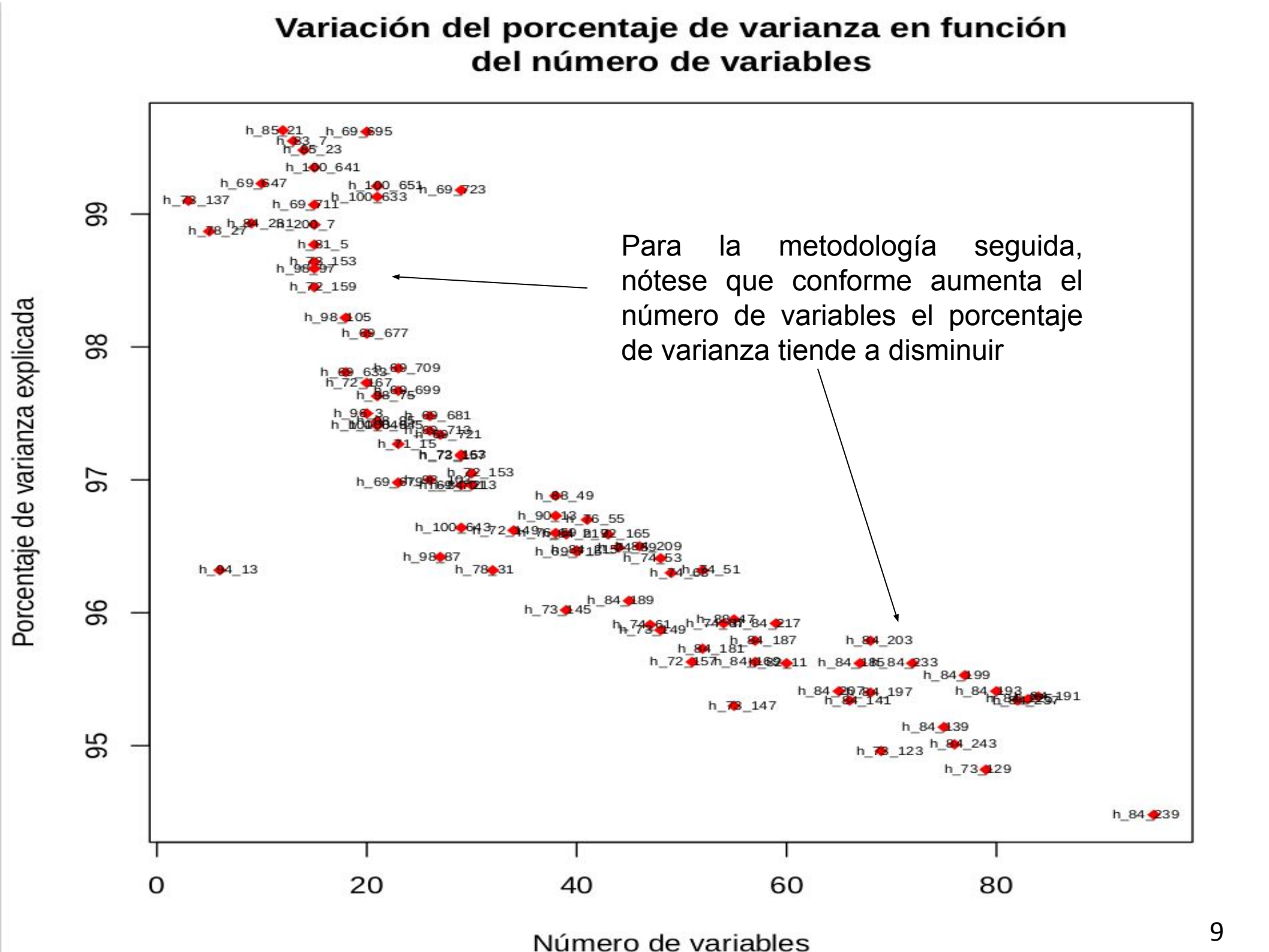
Entre mayor es el número de factores, mayor es la cantidad de varianzas espurias en la matriz de factores. Aunque a primera vista su contribución es pequeña, conjuntamente podrían contribuir significativamente a mejorar el porcentaje total de varianza explicada.



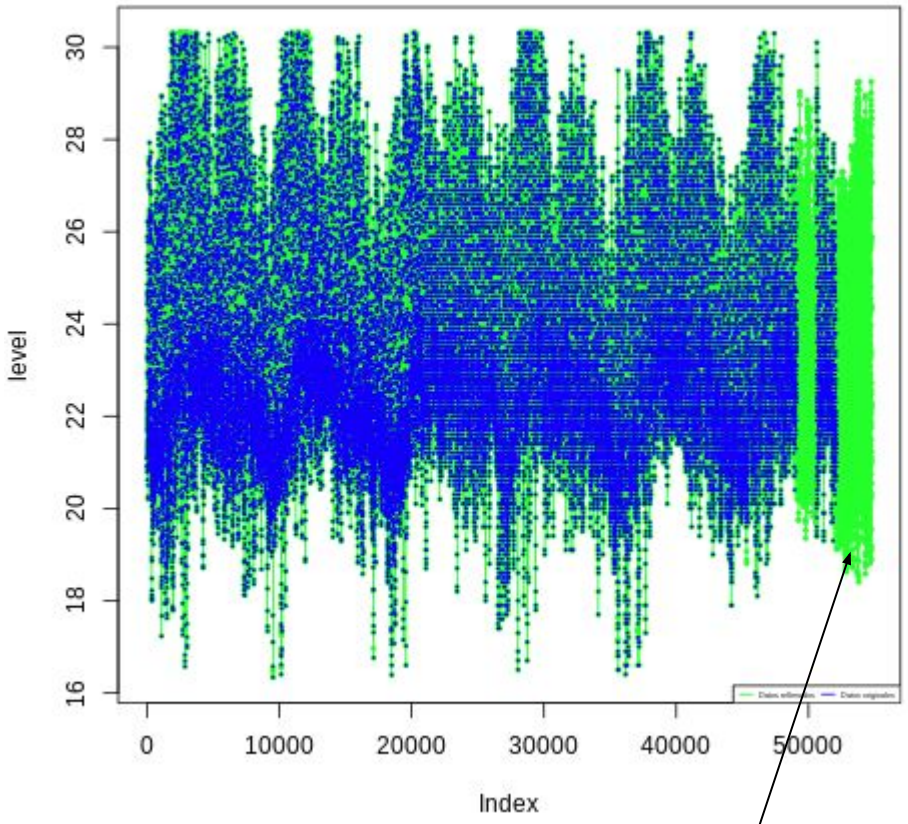


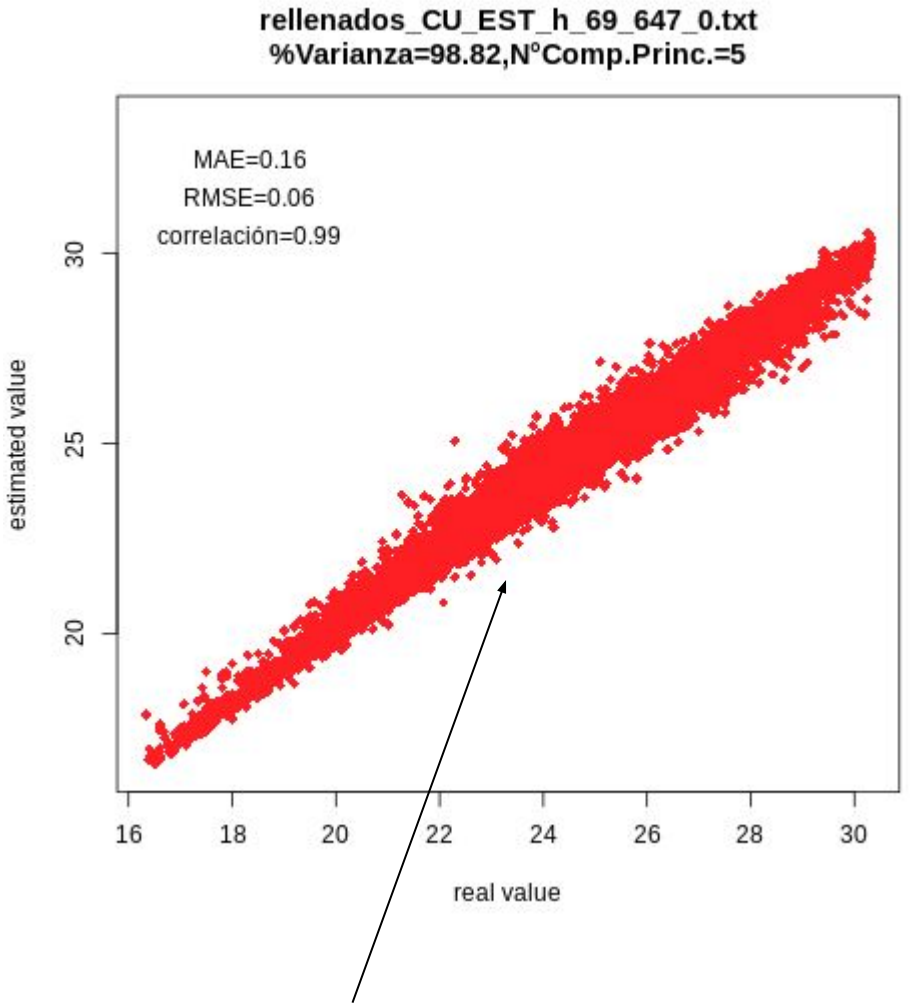



Porcentaje de varianza contra
número de componentes principales usadas
Porcentage de varianza
95
96
97
98
99
2
4
6
8
10
12
14
Número de Componentes principales


Variación del porcentaje de varianza en función del número de variables
Porcentaje de varianza explicada
Número de variables
99
98
97
96
95
0
20
40
60
80
Para la metodología seguida, nótese que conforme aumenta el número de variables el porcentaje de varianza tiende a disminuir

Resultados: Finca Brazilia

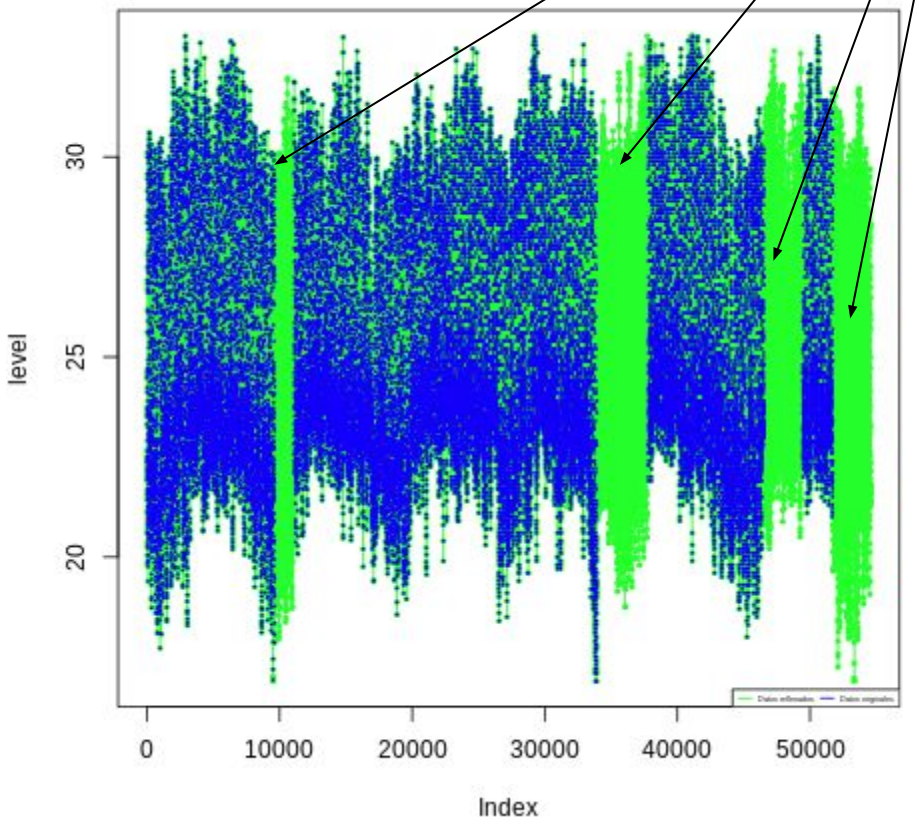


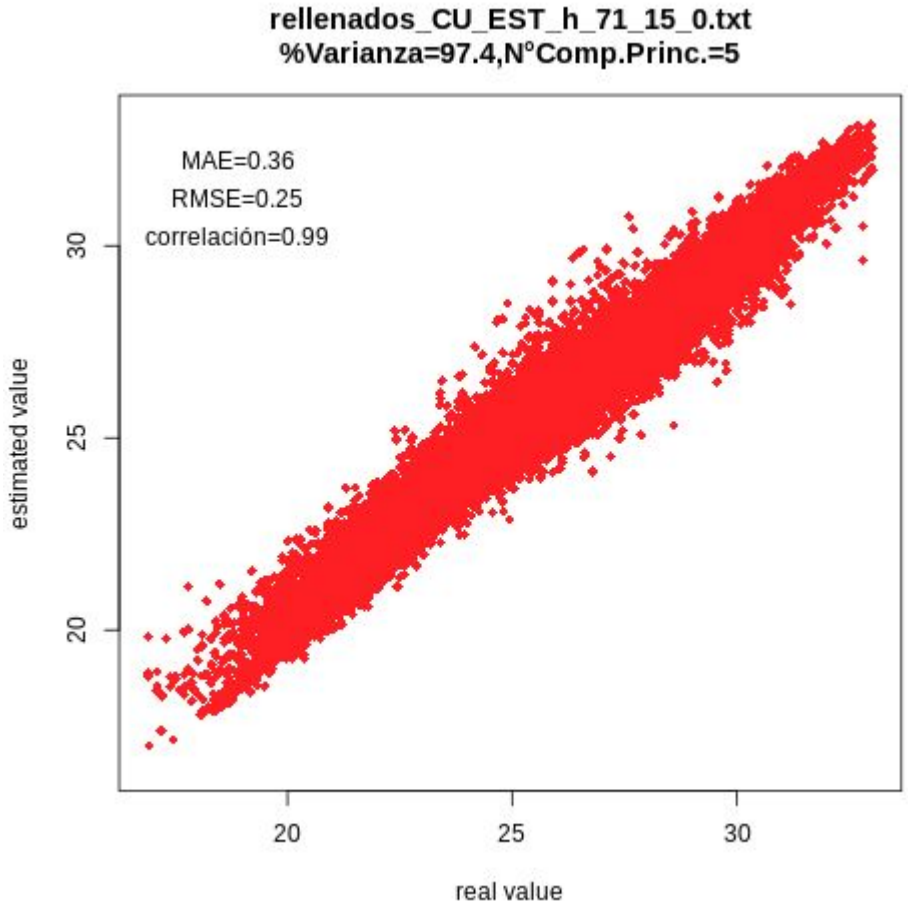


Estos datos son valores predichos por los datos de estaciones circunvecinas

¡Diagrama de dispersión muestra buena concordancia entre datos rellenados y datos reales!!!

Resultados: Cantagallo





¿Es ePCA mejor o no?

Relleno usando PCA ordinario 0.8<r²<0.9

Tabla 3: Errores por estación

Estación	MBE	MAE	RMSE
Pinilla	0.41	1.15	1.39
La Ceiba	0.35	1.38	1.63
Paquera	0.41	1.13	1.40
Santa Cruz	0.23	1.40	1.63
Mojica	0.05	1.21	1.43
Todas	0.29	1.25	1.49

Relleno usando PCA ordinario, r²=0.87

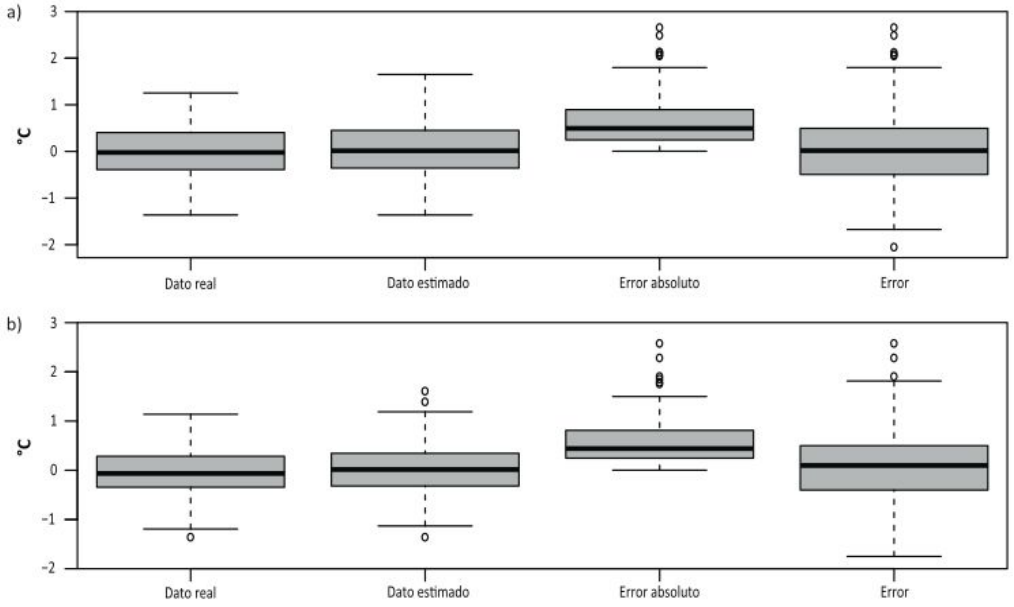


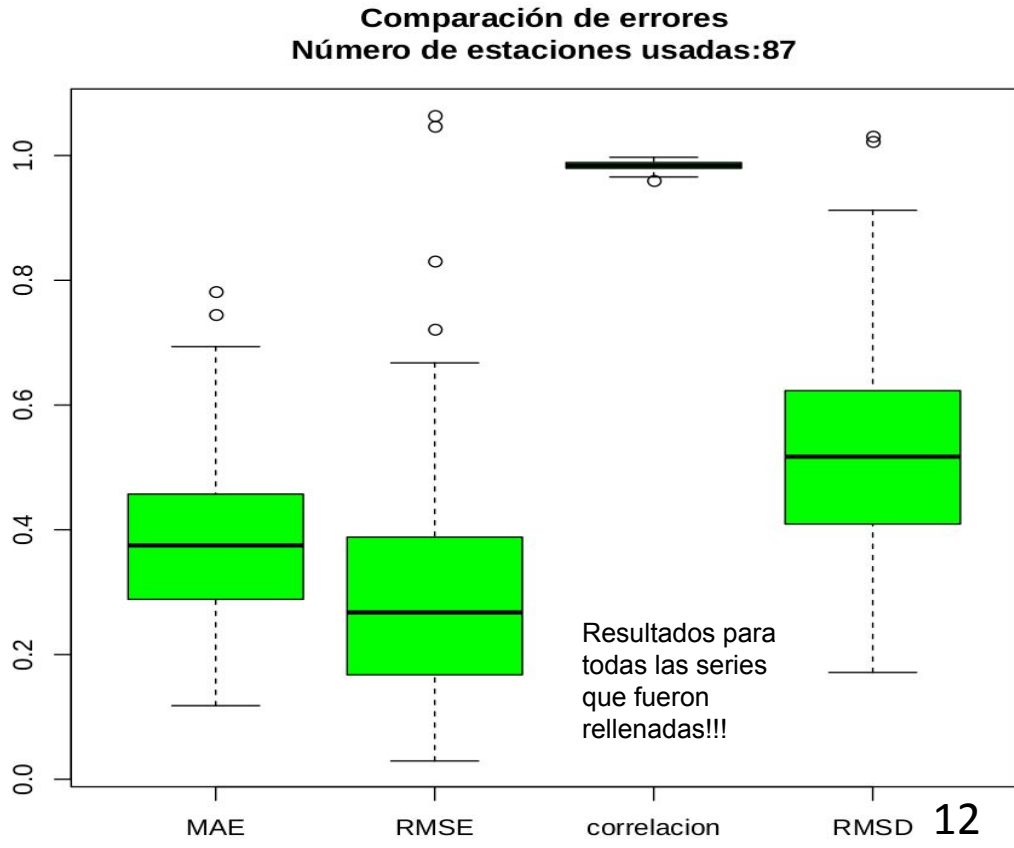
Figura 9. Gráfico de cajas y bigotes para los errores: a) Aeropuerto Internacional Juan Santamaría. b) Estación Experimental Fabio Baudrit.

Cuadro 3: Errores por estación

Estación	MBE	MAE	RMSE
Aeropuerto Internacional Juan Santamaría	0,04	0,63	0,81
Estación experimental Fabio Baudrit	0,07	0,56	0,70

Relleno usando ePCA

Estación	MAE
Pinilla	0.376
La Ceiba	0.526
Paquera	0.406
Santa Cruz	0.540
Hacienda Mojica	0.370



Resultados para todas las series que fueron rellenadas!!!

La parte triste...

Aproximadamente un 50% de las 172 estaciones analizadas no cumplen los criterios de selección para la metodología de relleno usada aquí.

En nómina	No existen en la nómina	series actuales con grupos	Grupo existe pero no hay serie	Series analizadas sin grupos	Series rellenadas	Sin rellenar
178	0	172	52	6	87	85

¿Estaciones de la red metropolitana?

85-52=33---> Incluye estaciones cerradas

Trabajos futuros

- Relleno de otros parámetros meteorológicos además de temperatura superficial horaria.
- Estudiar aspectos de eficiencia de diversos métodos.
- Evaluación de métodos de aprendizaje de máquina para el relleno de series temporales (mejorar ePCA).
- Evaluación de datos híbridos de acceso libre tales como CHIRPS y datos generados por modelos en la mejora de las estimaciones de relleno y control de calidad.

Trabajos futuros

- Generar una base de datos híbridos que pueda usarse para mejorar los métodos de control de calidad y relleno.
- Unificar diversos registros y fuentes de eventos extremos para integrarlo en los procesos de revisión de los datos de una forma automática o semi-automática.
- Definir criterios con respecto al sitio de almacenamiento de los resultados, formatos y necesidades de otros departamentos en cuanto a datos con control de calidad, relleno y homogeneización.

Observaciones

- Los datos rellenados se van a ir agregando en el compartido “BDEspejo” en el servidor de respaldo de la Unidad de Informática del IMN.
- Otros parámetros meteorológicos y/o mejoras de los actuales se irán agregando conforme se avance con este trabajo.
- Se deben de buscar soluciones para los casos donde el uso de estaciones cercanas hace difícil la aplicación de los criterios de rellenado (Fuentes de datos alternativas).

¡Muchas gracias!