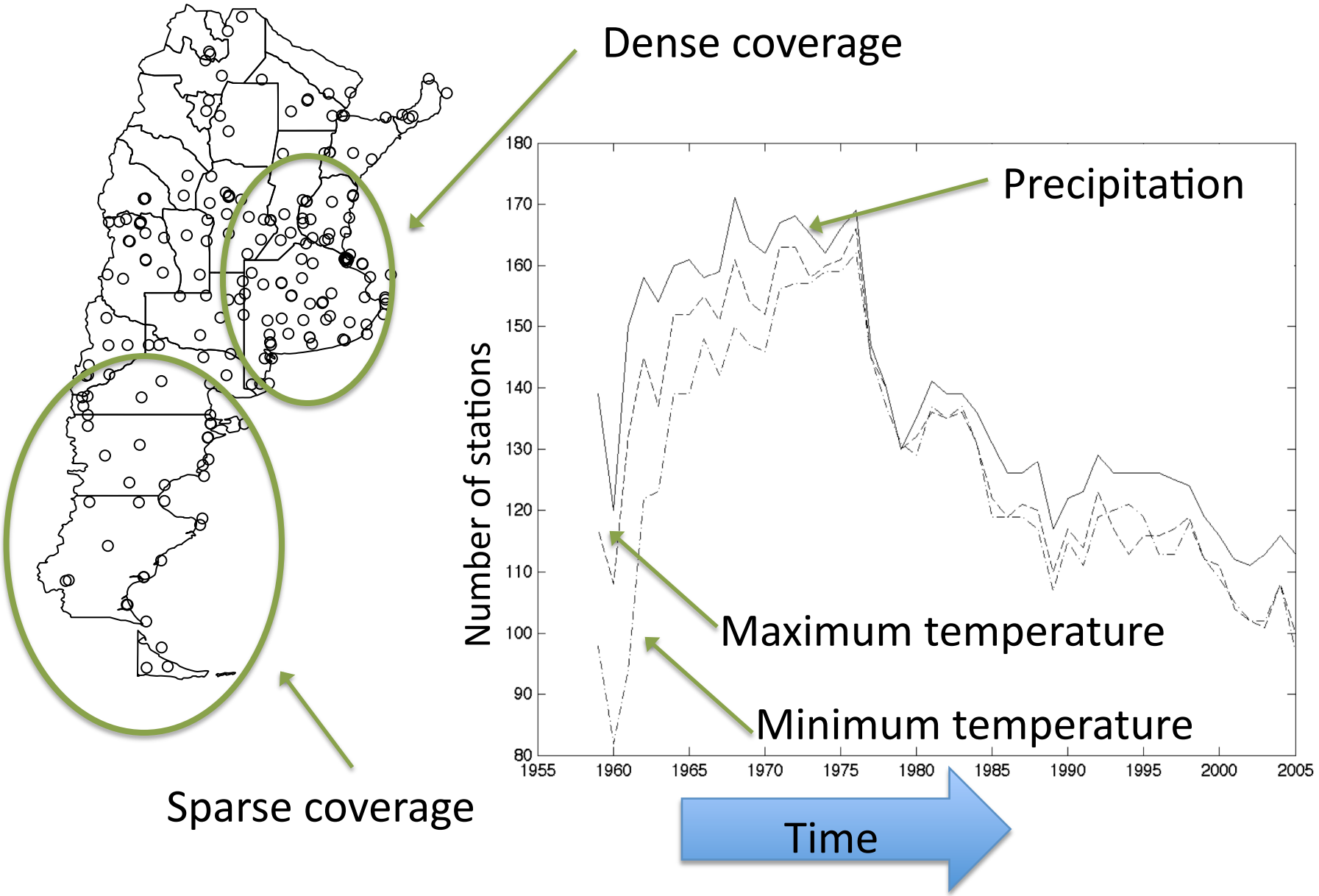
Error and Inhomogeneity detection in daily climate time series

Andrés Farall
Jean-phillipe Boulanger
Liliana Orellana

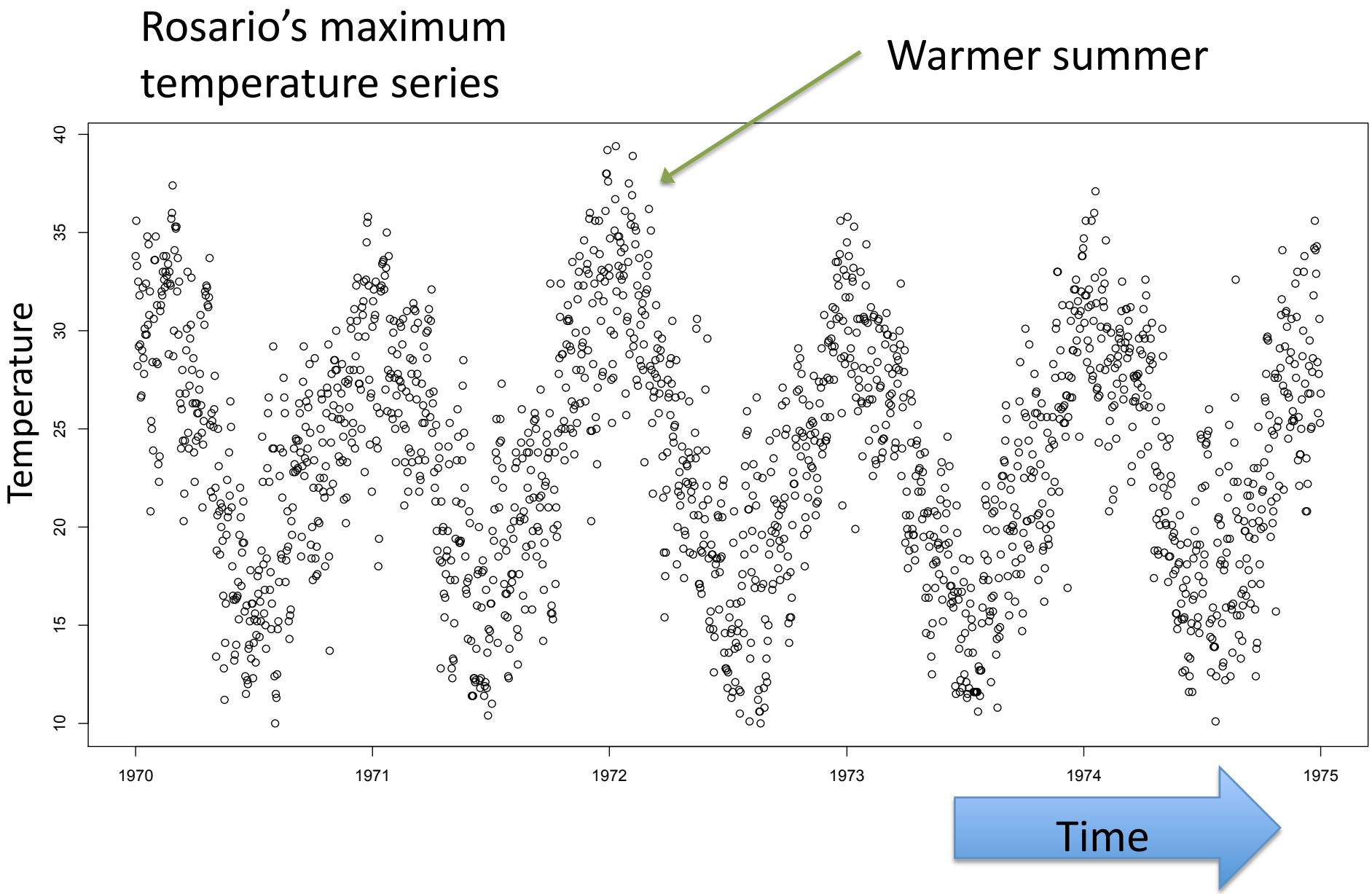
Goals

- To develop an error/inhomogeneity detection algorithm for meteorological data.
- To implement the proposed methods in the free open-source (GNU) language R.
- To apply the methods to the CLARIS LPB databases.

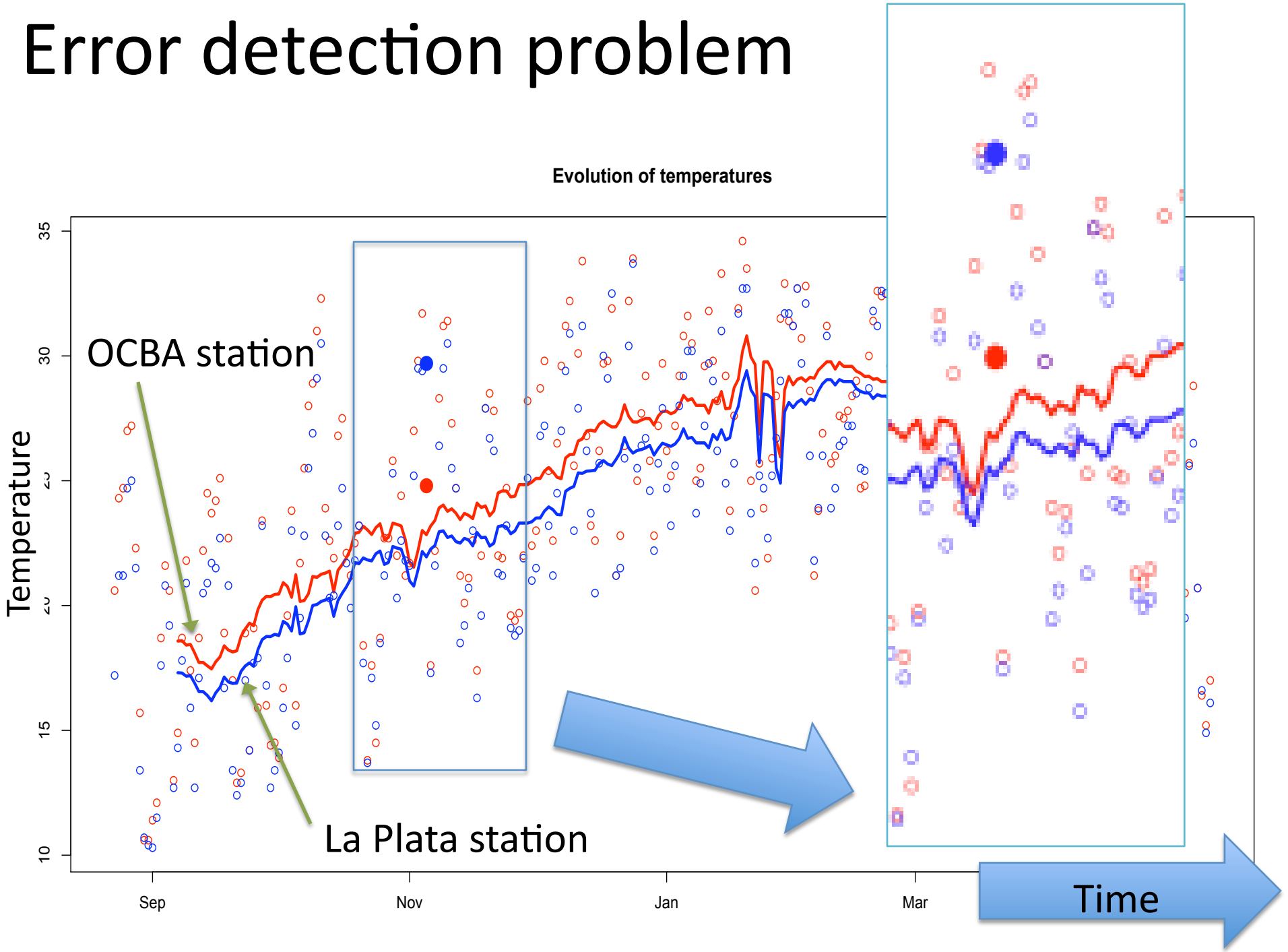
The monitoring stations



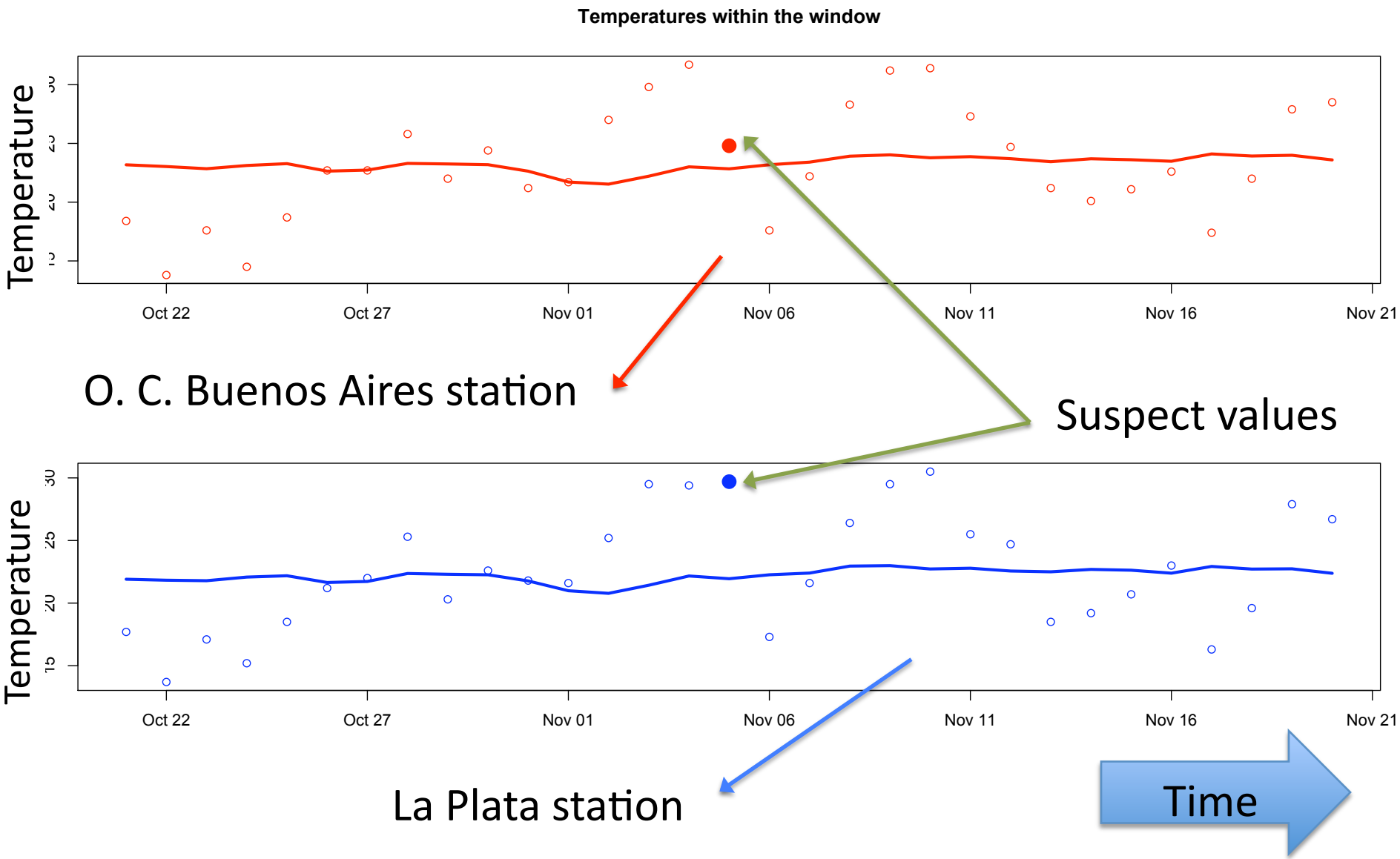
Temperature (maximum) data



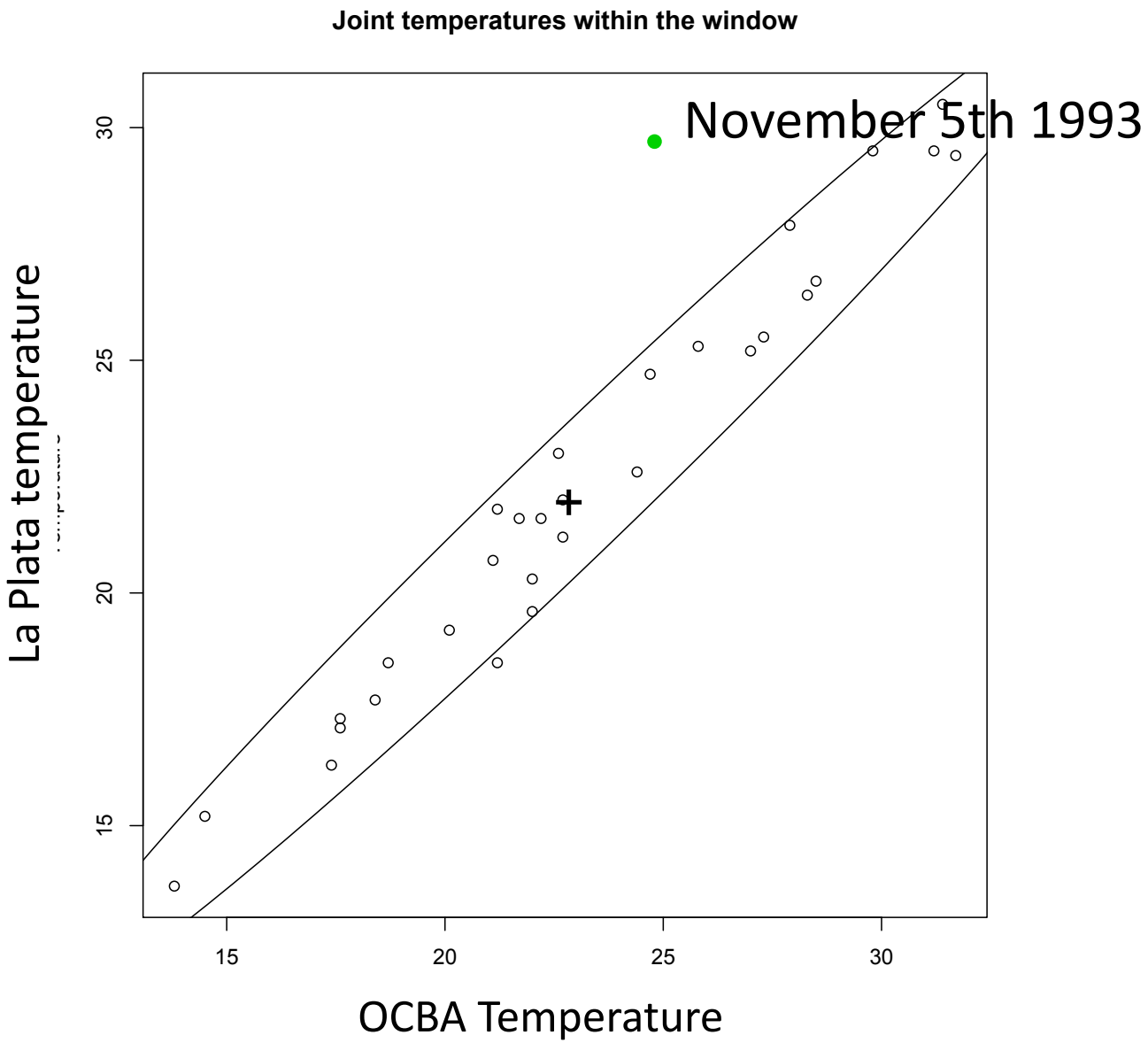
Error detection problem



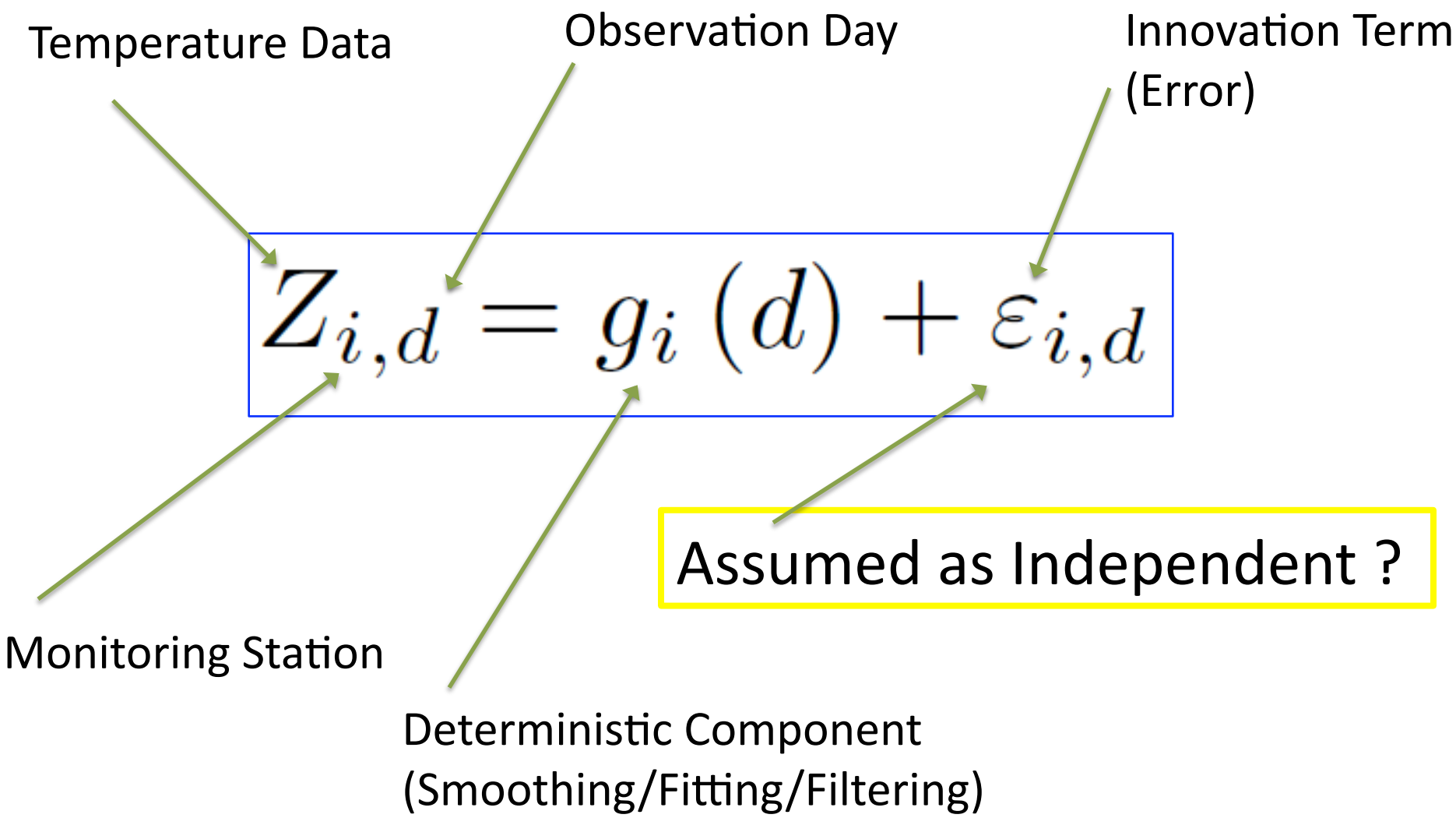
The temperature series



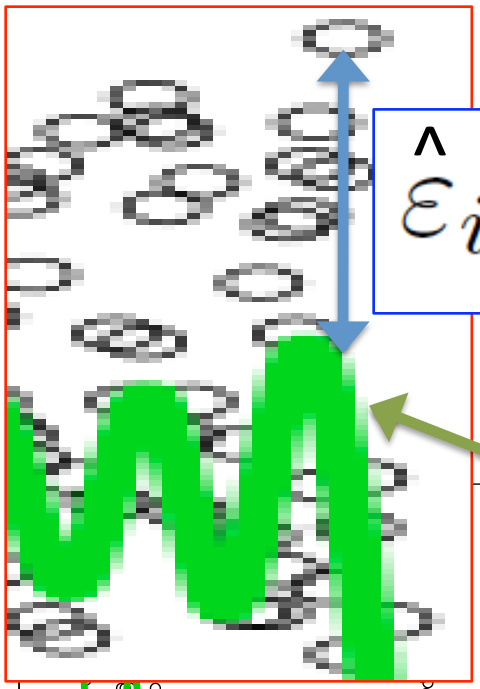
Joint temperature values



The setup


Temperature Data
Observation Day
Innovation Term
(Error)
$Z_{i,d} = g_i(d) + \varepsilon_{i,d}$
Assumed as Independent ?
Monitoring Station
Deterministic Component
(Smoothing/Fitting/Filtering)

Smoothing/Fitting/ Filtering



$\hat{\varepsilon}_{i,d}$

$\hat{g}_i\,(d)$

Rosario's maximum temperature series

Temperature

10 15 20 35 40

1970 1971 1972 1973 1974 1975

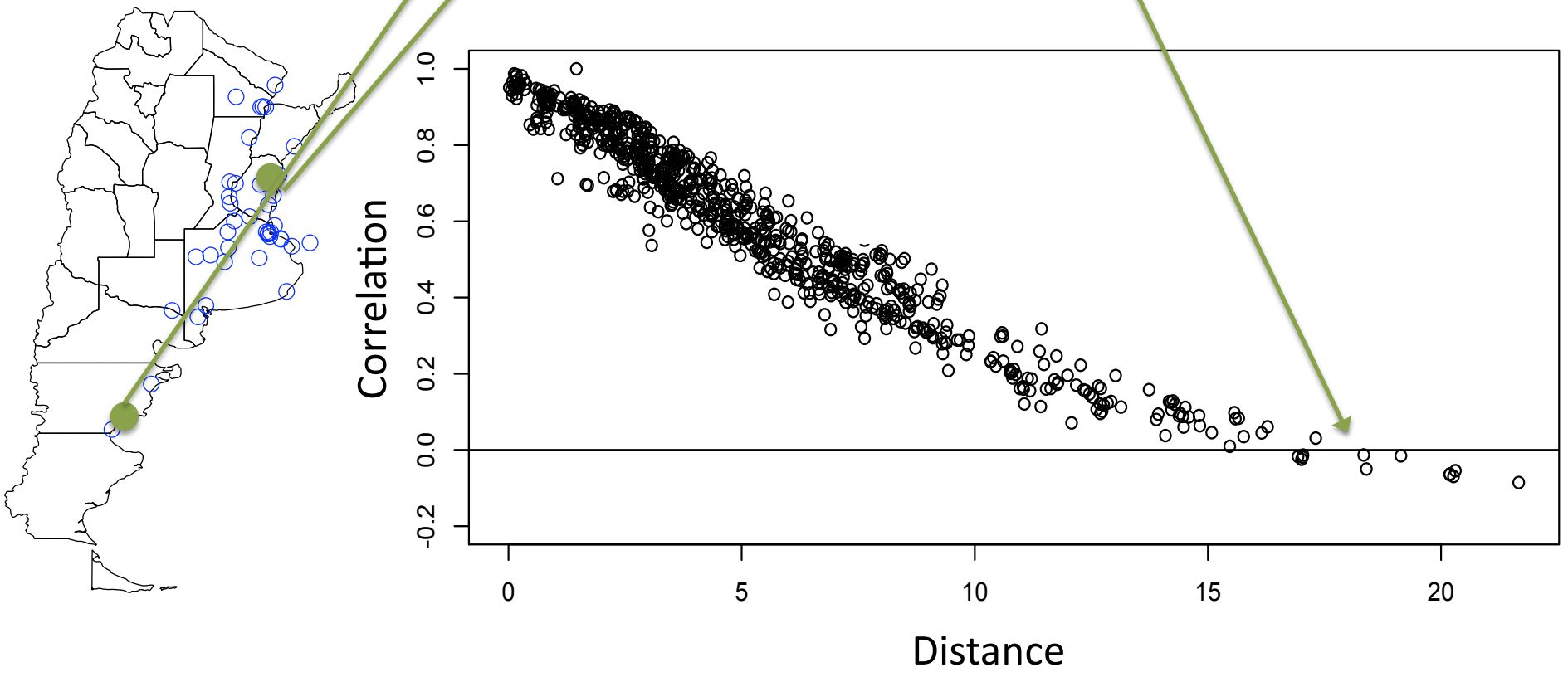
Time

Smoothing/Fitting/Filtering Choices

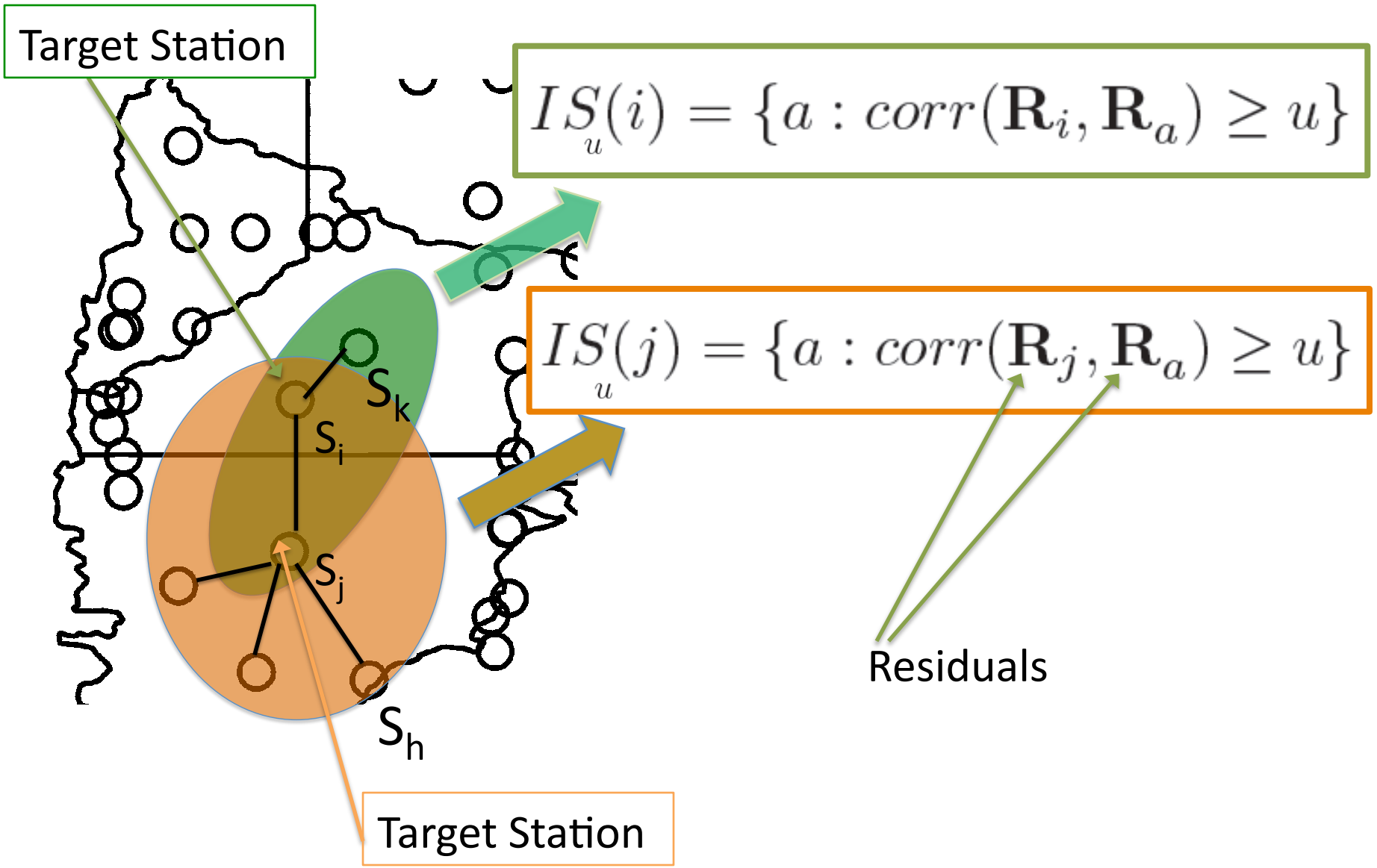
- Running mean
- Running median
- LOWESS
- Deseasonalization.
- ARIMA/GATCH fitting.
- Fast Fourier Transform (FFT)
- Filtering.

Historical Correlation Matrix of Residuals

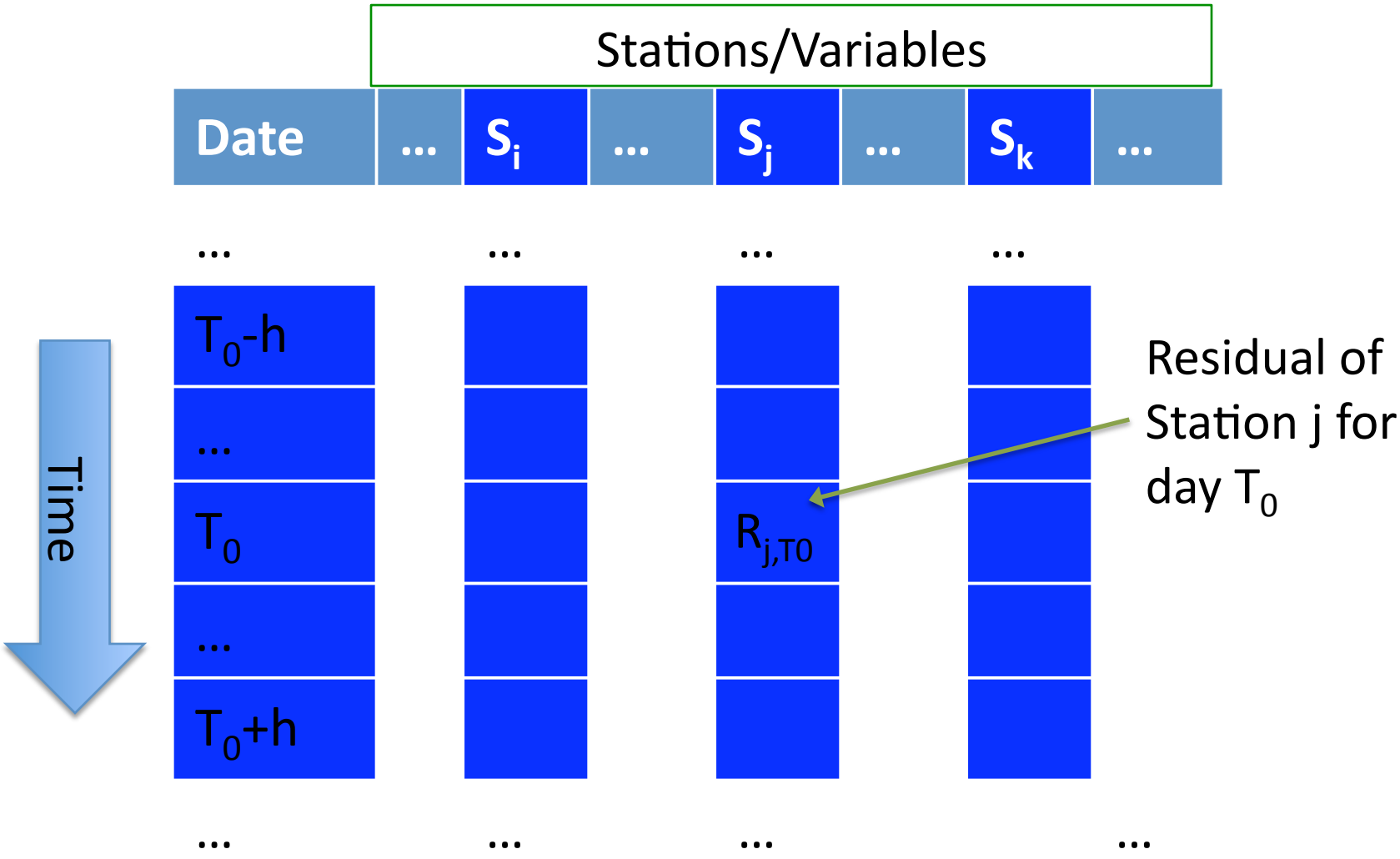
		Comodoro	Concordia	
Comodoro		1	0.01	
Concordia		0.01	1	



Influence Set of a Target Station



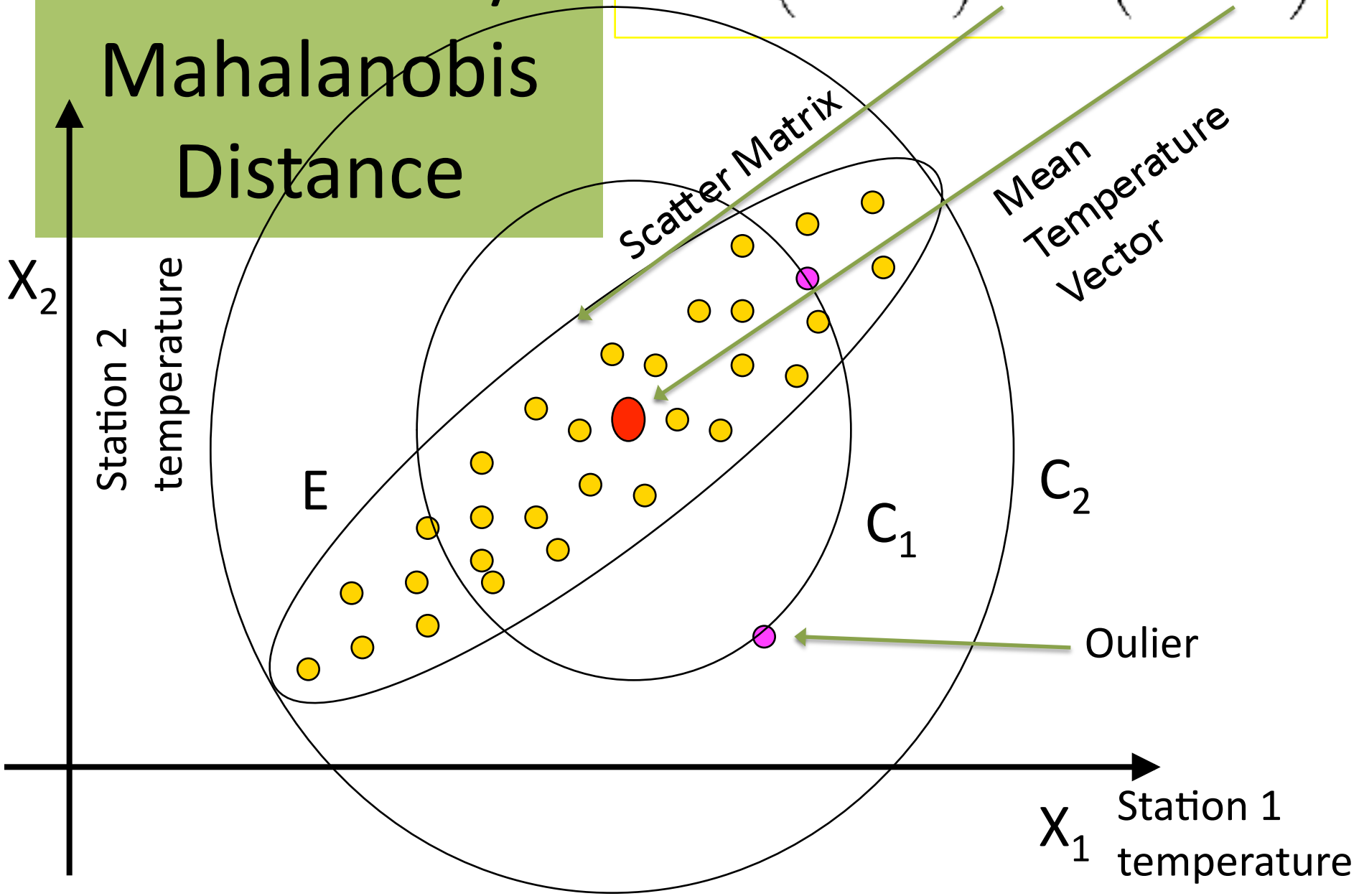
So, for each Influence Set we have:



What is an Outlying Multivariate Observation?

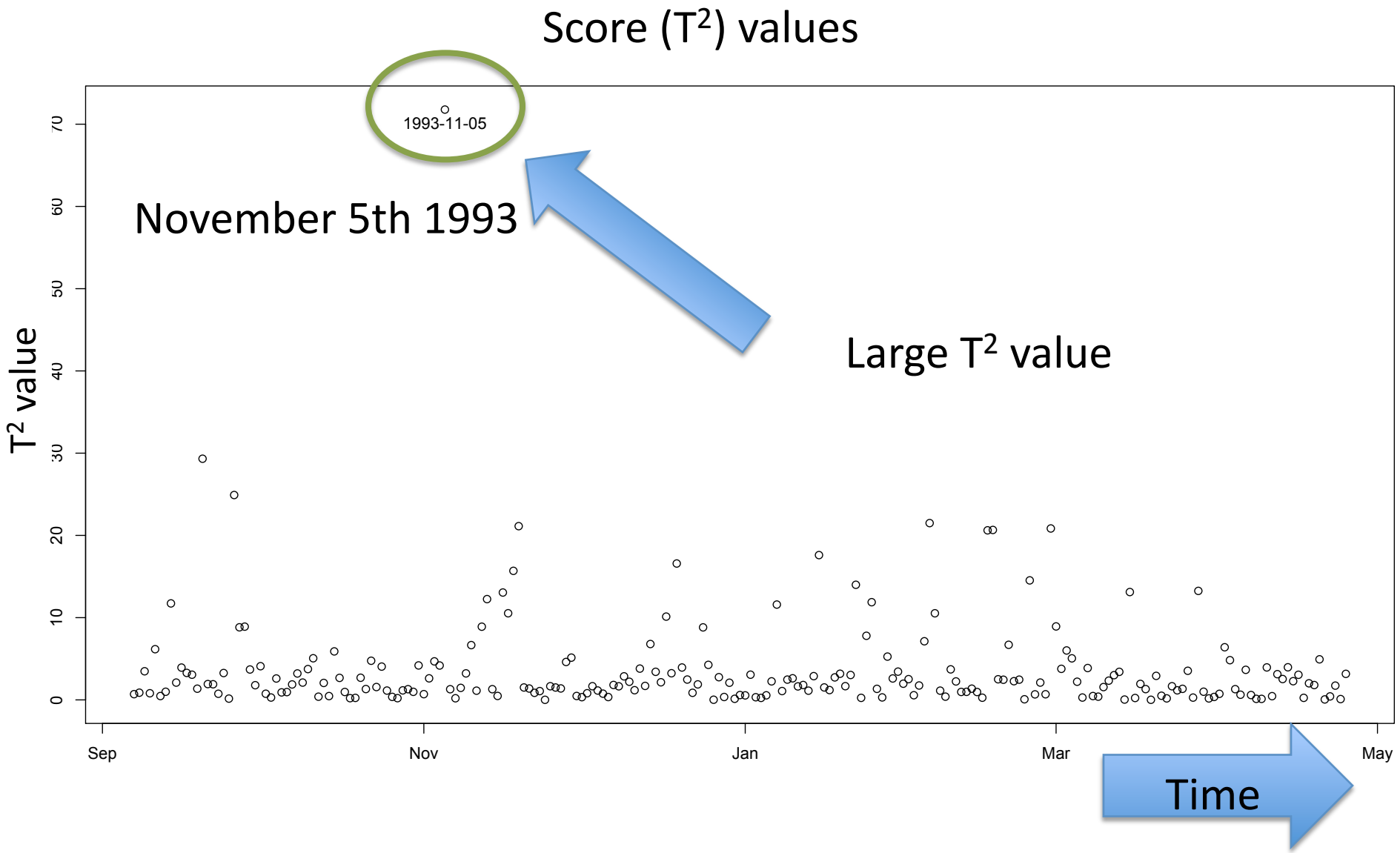
T^2 Statistic / Mahalanobis Distance

$$T^2 = \left(\mathbf{X} - \bar{\mathbf{X}}\right)' \mathbf{S}^{-1} \left(\mathbf{X} - \bar{\mathbf{X}}\right)$$

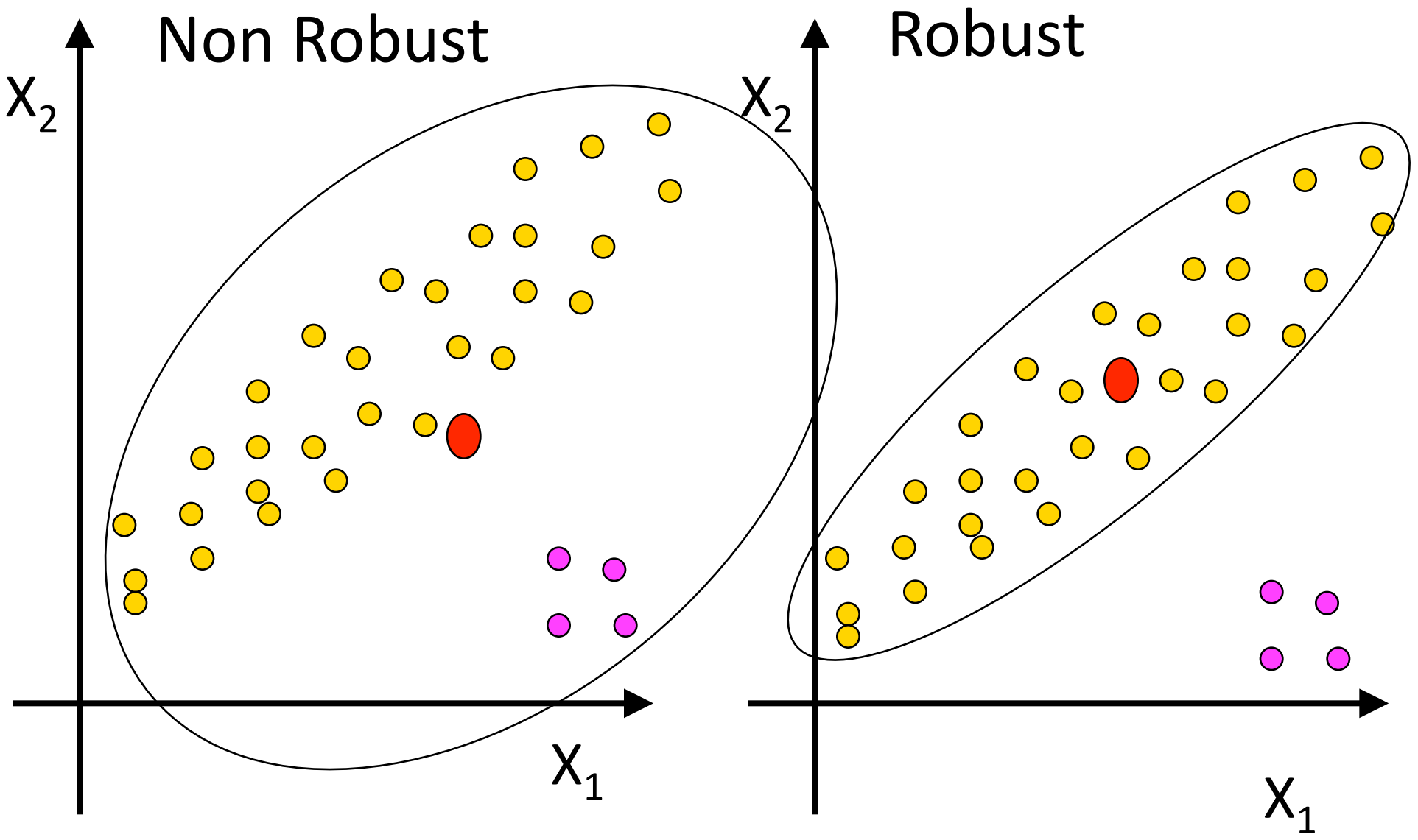


Measuring Outlyingness :

La Plata and Buenos Aires



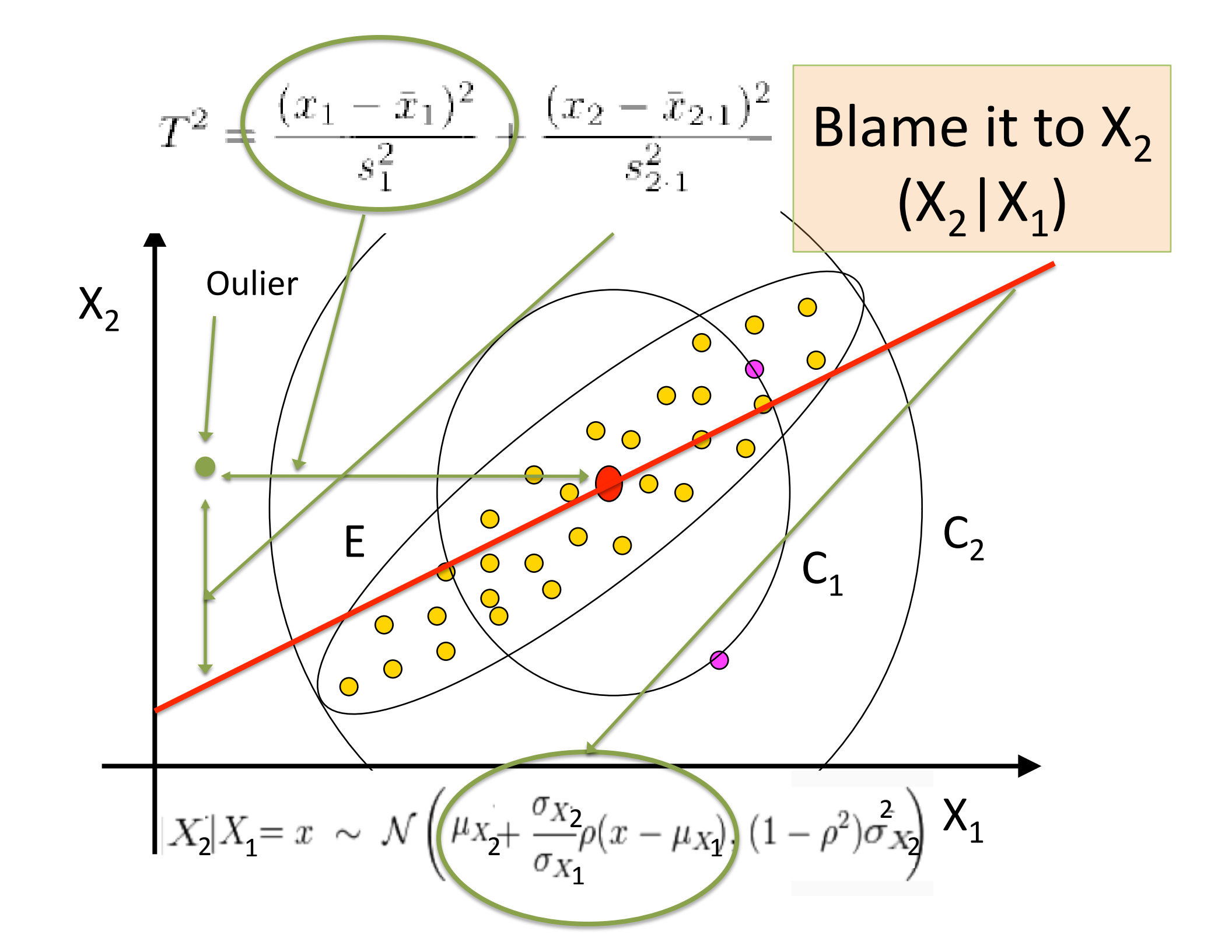
The Need for Robustness

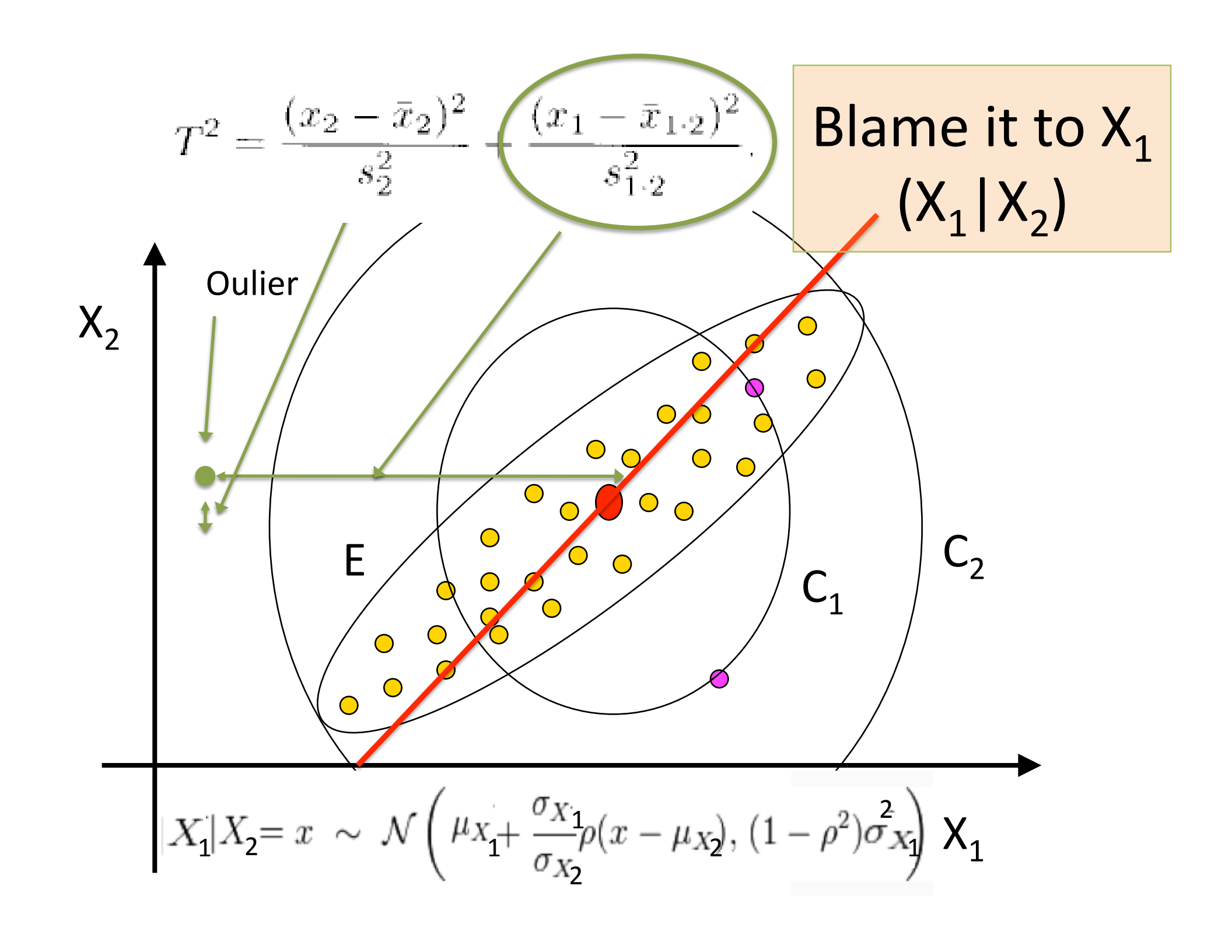


Local Time Span Selection

- A local time span determination is needed to compute the T^2 statistic. This value controls the number of days used to compute the covariance matrix and the location vector.
- If the span is too long then the “local (time) effect” could be lost.
- If the span is too short then the method yields unstable estimations (# obs >> # var).

Is There a Unique Station/Variable Responsible for the Large T^2?


$T^2 = \frac{(x_1 - \bar{x}_1)^2}{s_1^2} + \frac{(x_2 - \bar{x}_{2\cdot 1})^2}{s_{2\cdot 1}^2}$
Blame it to X_2 $(X_2|X_1)$
X_2
Oulier
E
C_1
C_2
X_1
$[X_2|X_1 = x \sim \mathcal{N}\left(\mu_{X_2} + \frac{\sigma_{X_2}}{\sigma_{X_1}}\rho(x - \mu_{X_1}), (1 - \rho^2)\sigma_{X_2}^2\right)$


$T^2 = \frac{(x_2 - \bar{x}_2)^2}{s_2^2} - \frac{(x_1 - \bar{x}_{1\cdot 2})^2}{s_{1\cdot 2}^2}$
Blame it to X_1
$(X_1|X_2)$
Oulier
X_2
E
C_1
C_2
X_1
$|X_1|X_2 = x \sim \mathcal{N}\left(\mu_{X_1} + \frac{\sigma_{X_1}}{\sigma_{X_2}}\rho(x - \mu_{X_2}), (1 - \rho^2)\sigma_{X_1}^2\right)$

MYT Decomposition

$$T^2 = T_1^2 + T_{2/1}^2 + T_{3/1,2}^2 + T_{4/1,2,3}^2 + \cdots + T_{p/1,2,3,4,\ldots,p-1}^2$$

$$T_1^2 = \frac{(x_1 - \overline{x}_1^2)^2}{S_1^2}$$

$$T_{p/1,2,3,4,\ldots,p-1}^2 = \frac{(x_p - \overline{x}_{p/1,2,3,4,\ldots,p-1}^2)^2}{S_{p/1,2,3,4,\cdots,p-1}^2}$$

Suspect station

The Distance Contribution

Let T^2 be the Mahalanobis distance, the MYT decomposition is

$$T^2 = T_1^2 + T_{2/1}^2 + T_{3/1,2}^2 + T_{4/1,2,3}^2 + \ldots + T_{p/1,2,3,4,\ldots,p-1}^2$$

We define the contribution of variable p to the distance as

$$C_p = \frac{T_{p/1,2,3,4,\ldots,p-1}^2}{T^2}$$

because all terms are positive, the contribution satisfies

$$0 \leq C_p \leq 1$$

Choosing *p* decomposiitions (between *p*! distinct decompositions)

$$T^2 = T^2_2 + T^2_{3/2} + T^2_{4/2,3} + T^2_{5/2,3,4} + \ldots + T^2_{1/2,3,4,\ldots,p}$$

$$T^2 = T^2_1 + T^2_{3/1} + T^2_{4/1,3} + T^2_{5/1,3,4} + \ldots + T^2_{2/1,3,4,\ldots,p}$$

$$T^2 = T^2_1 + T^2_{2/1} + T^2_{4/1,2} + T^2_{5/1,2,4} + \ldots + T^2_{3/1,2,4,\ldots,p}$$

$$\vdots \qquad\qquad \vdots \qquad\qquad \vdots$$

$$T^2 = T^2_1 + T^2_{2/1} + T^2_{3/1,2} + T^2_{4/1,2,3} + \cdots + T^2_{p-1/1,2,3,4,\ldots,p-2,p}$$

$$T^2 = T^2_1 + T^2_{2/1} + T^2_{3/1,2} + T^2_{4/1,2,3} + \ldots + T^2_{p/1,2,3,4,\ldots,p-1}$$

Error Detection

- If in a certain day the following conditions happen
  - T^2 is large
  - $C_p = \frac{T^2_{p/1,2,3,4,\ldots,p-1}}{T^2}$ is close to 1

  then there is an error in station *p*,

  because all other stations agree

T^2_{j/A_j} Is small

for all $j \in \{1, 2, 3, 4, \ldots, p-1\}$ and $A_j \subseteq \{1, 2, 3, 4, \ldots, j-1, j+1, \ldots, p-1\}$

Robust Location-Scatter Estimators

- Minimum Covariance Determinant (MCD).
- Minimum Volume Elipsoid (MVE).
- Stahel-Donoho Estimator.
- Orthogonalized Gnanadesikan/Kettenring (OGK).
- Fixed Correlation Scatter Estimator

The Current Implementation

- MCD → Slow
- OGK → Fast
- Fixed Correlation Scatter Estimate → Superfast

The MCD Estimator

Given a *p* dimensional data set $\mathbf{X}=\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$,

the **MCD** estimator (Rousseeuw, 1984) is defined as the subset of *h* observations out of *n* whose classical covariance matrix has the smallest determinant.

The MCD location estimator **T** is defined as the mean of that subset the MCD scatter estimator.

The OGK Estimator

- The Orthogonalized Gnanadesikan/Kettenring estimator (Maronna and Zammar 2002), is a relatively fast scatter estimator based on the robust pairwise estimation of bivariate variances.

The Fixed Correlation Scatter Estimator

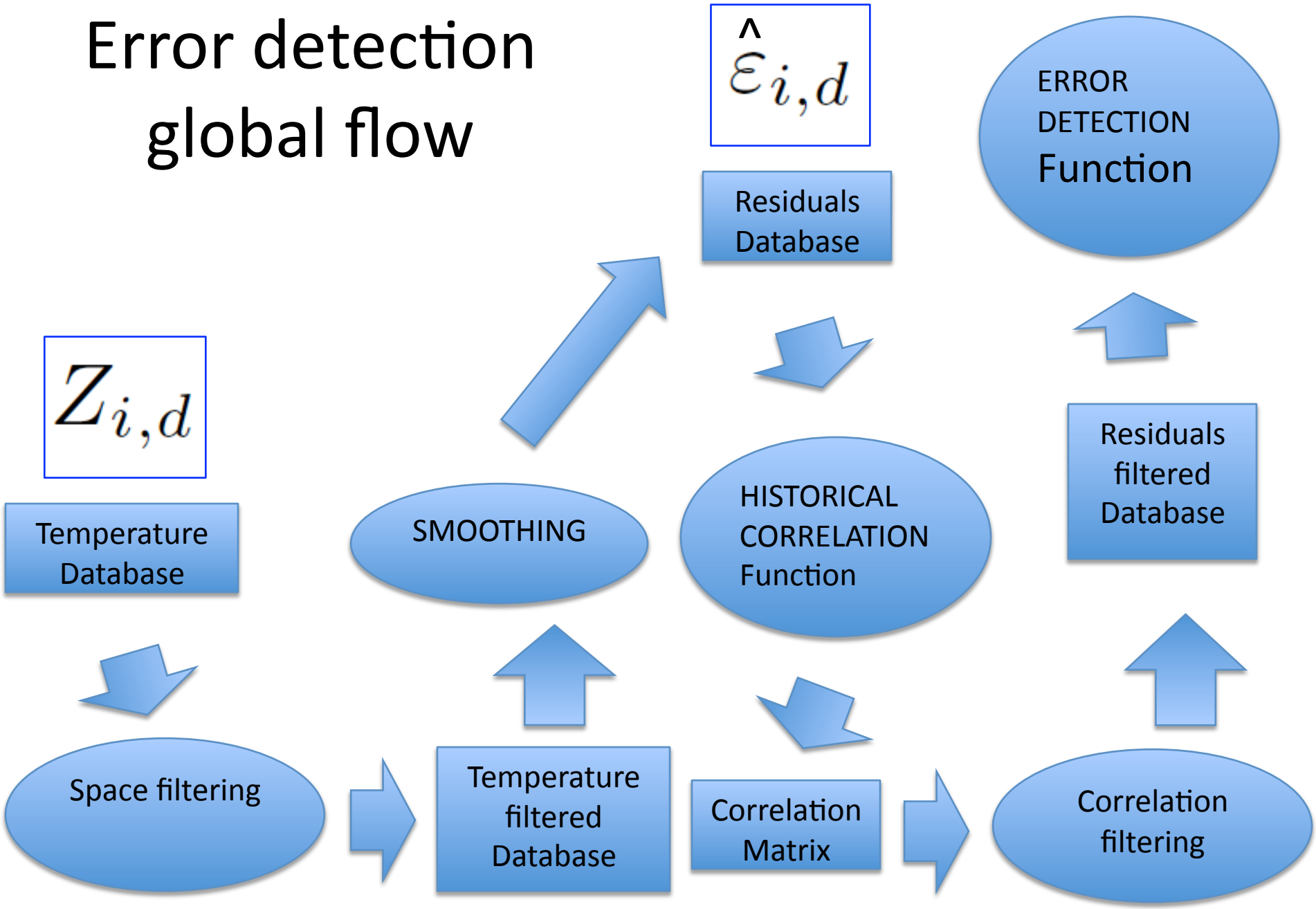
- This estimator relies on the fact that the correlation between monitoring stations do not change over time.

$$\sigma_{ij}(t) = \rho_{ij}\, \sigma_i(t)\, \sigma_j(t)$$

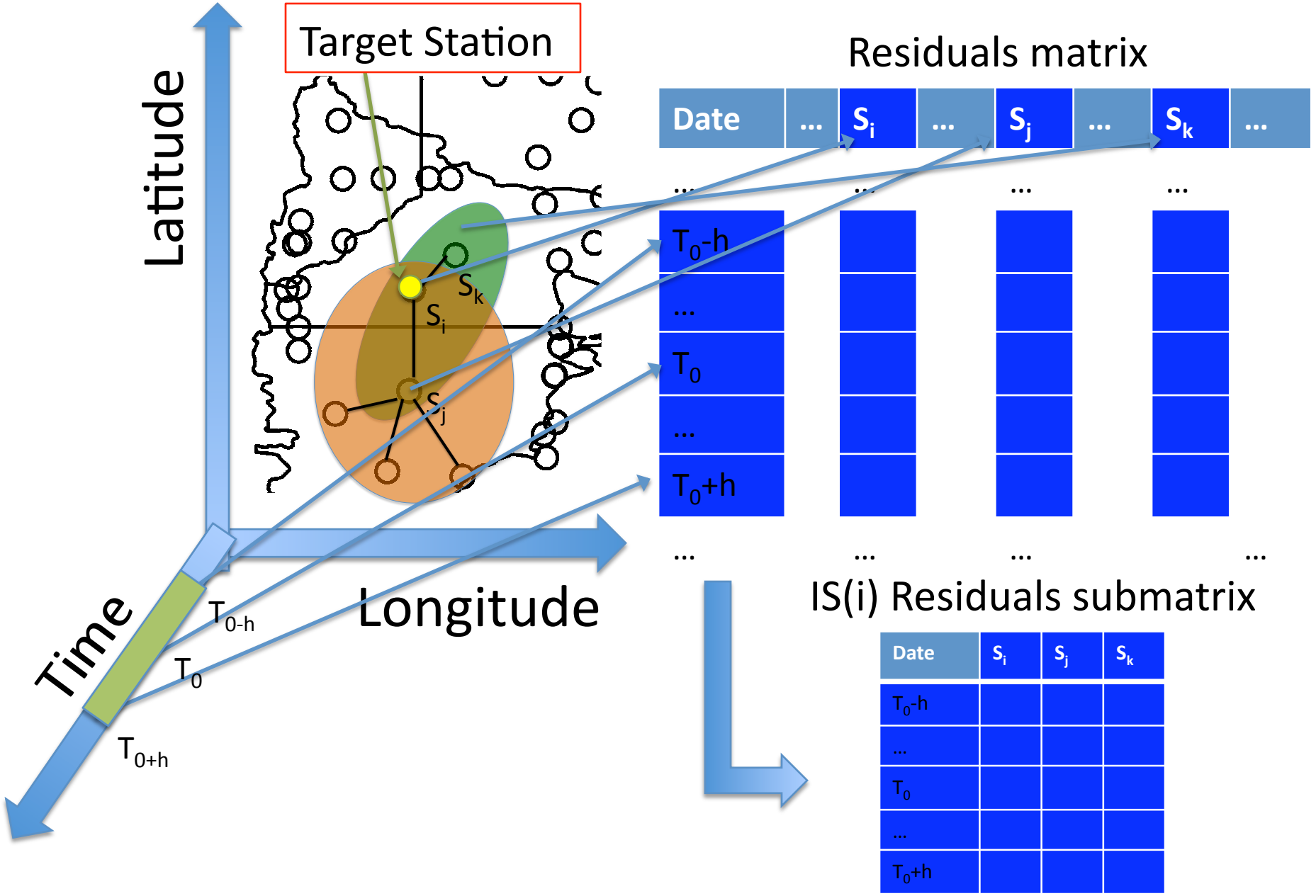
- Thus, to estimate the covariance, only robust estimations of univariate variances are needed.

The Method at Work

Error detection global flow



Error Detection method



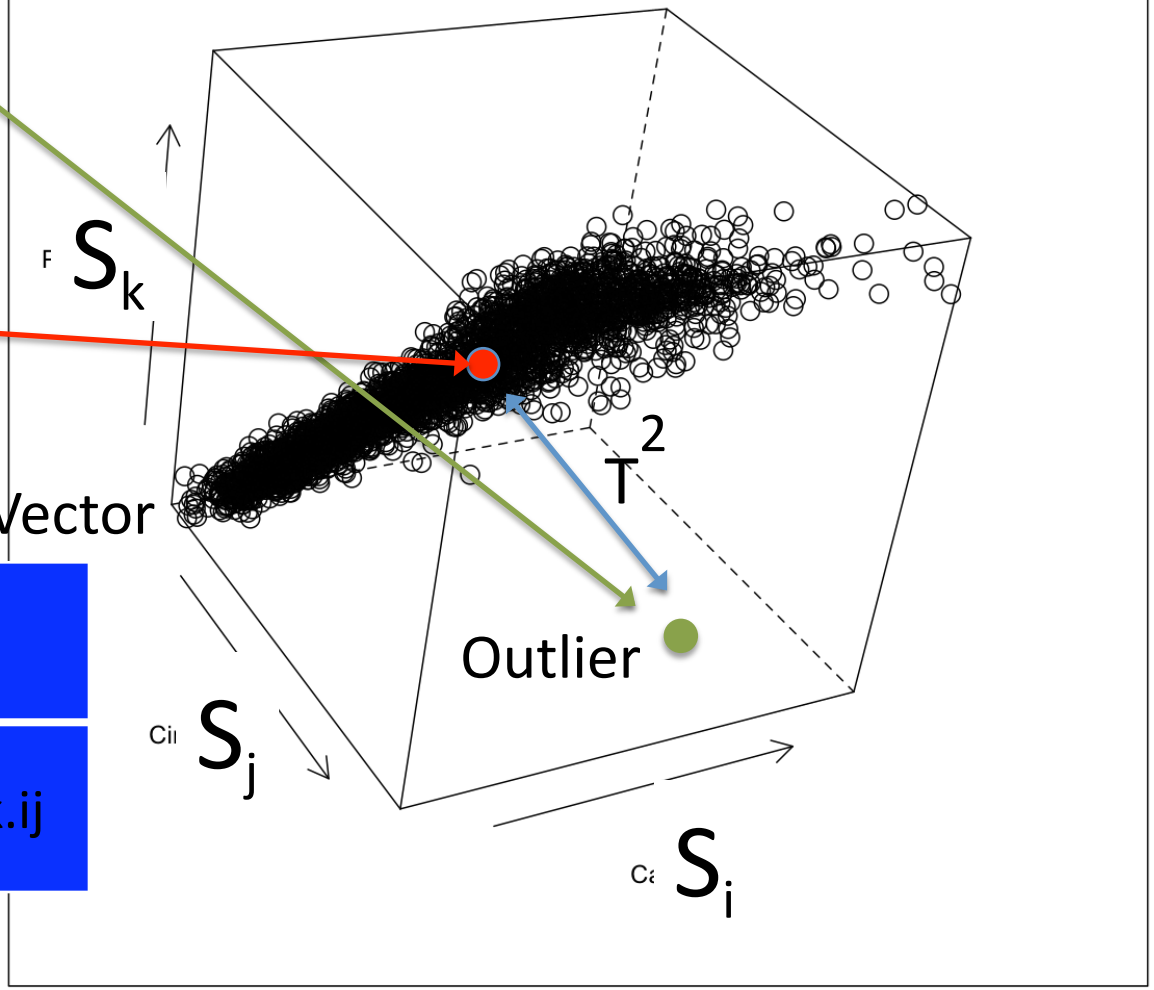
Error Detection method

IS(i) Residuals submatrix

Date	S_i	S_j	S_k
T_0-h			
...			
T_0			
...			
T_0+h			
Mean			

Time

Residuals scatter plot of stations I, j and k

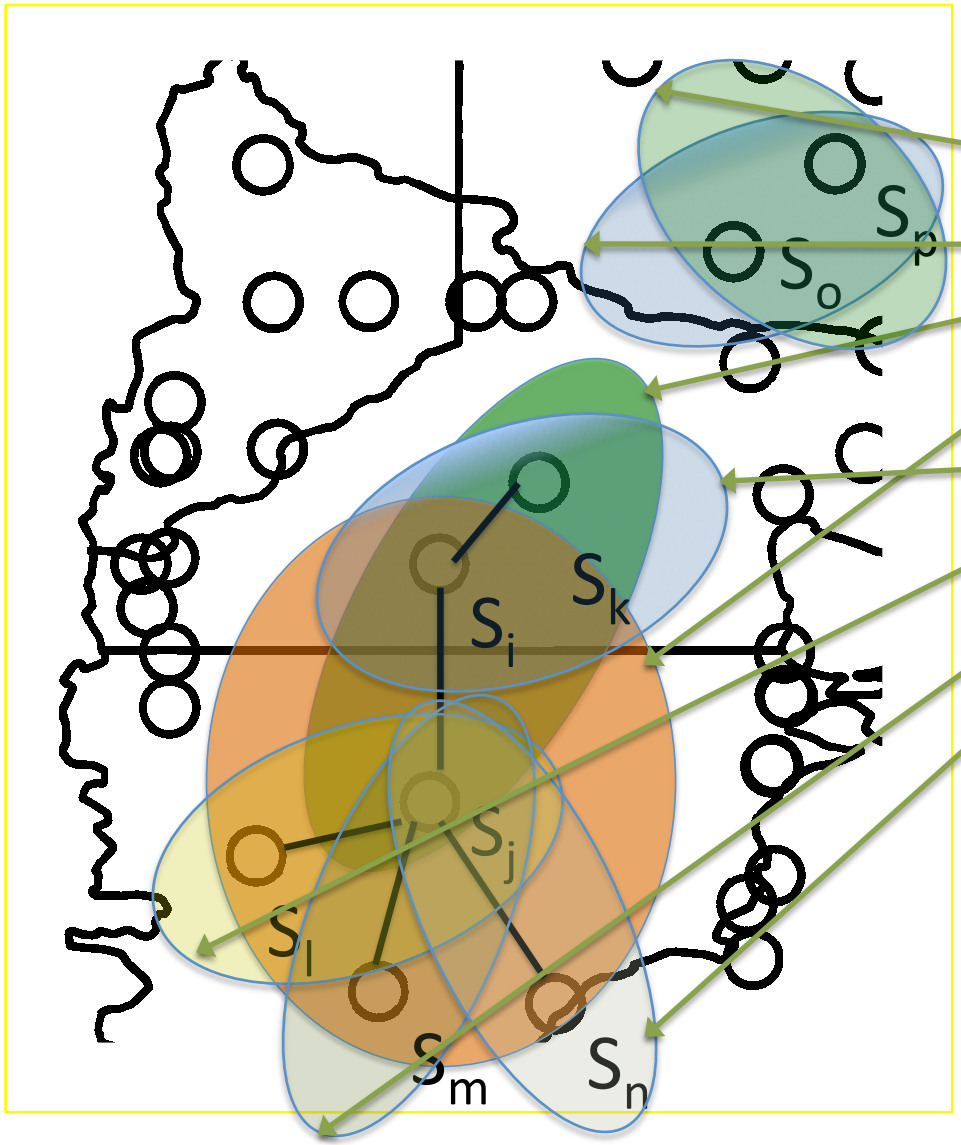


IS(i) Distance Contribution Vector

Date	S_i	S_j	S_k
T_0	$T^2_{i.jk}$	$T^2_{j.ik}$	$T^2_{k.ij}$

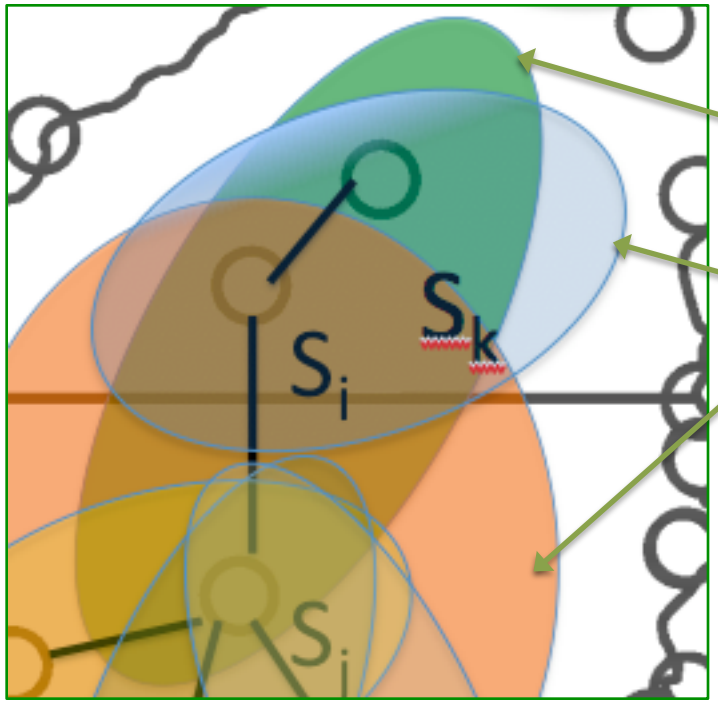
How to Summarize the Outlyingness Evidence from Different Influence Sets ?

A Toy Example (8 stations)



$$IS_o = \{S_o, S_p\}$$
$$IS_p = \{S_p, S_o\}$$
$$IS_i = \{S_i, S_j, S_k\}$$
$$IS_j = \{S_i, S_j, S_l, S_m, S_n\}$$
$$IS_k = \{S_i, S_k\}$$
$$IS_l = \{S_j, S_l\}$$
$$IS_m = \{S_j, S_m\}$$
$$IS_n = \{S_j, S_n\}$$

An Error Occurs in S_i



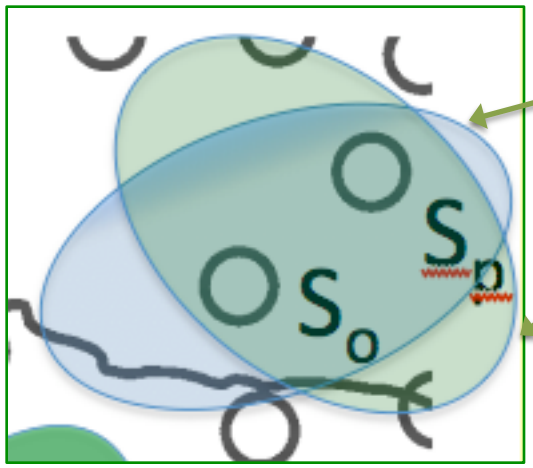
$$IS_o = \{S_o, S_p\}$$
$$IS_p = \{S_p, S_o\}$$
$$IS_i = \{S_i, S_j, S_k\}$$
$$IS_j = \{S_i, S_j, S_l, S_m, S_n\}$$
$$IS_k = \{S_i, S_k\}$$
$$IS_l = \{S_j, S_l\}$$
$$IS_m = \{S_j, S_m\}$$
$$IS_n = \{S_j, S_n\}$$

The three Influence sets are distinct

Three different distances are relevant here

$$T^2_{i/j,k}$$
$$T^2_{i/j,l,m,n}$$
$$T^2_{i/k}$$

An Error Occurs in S_o



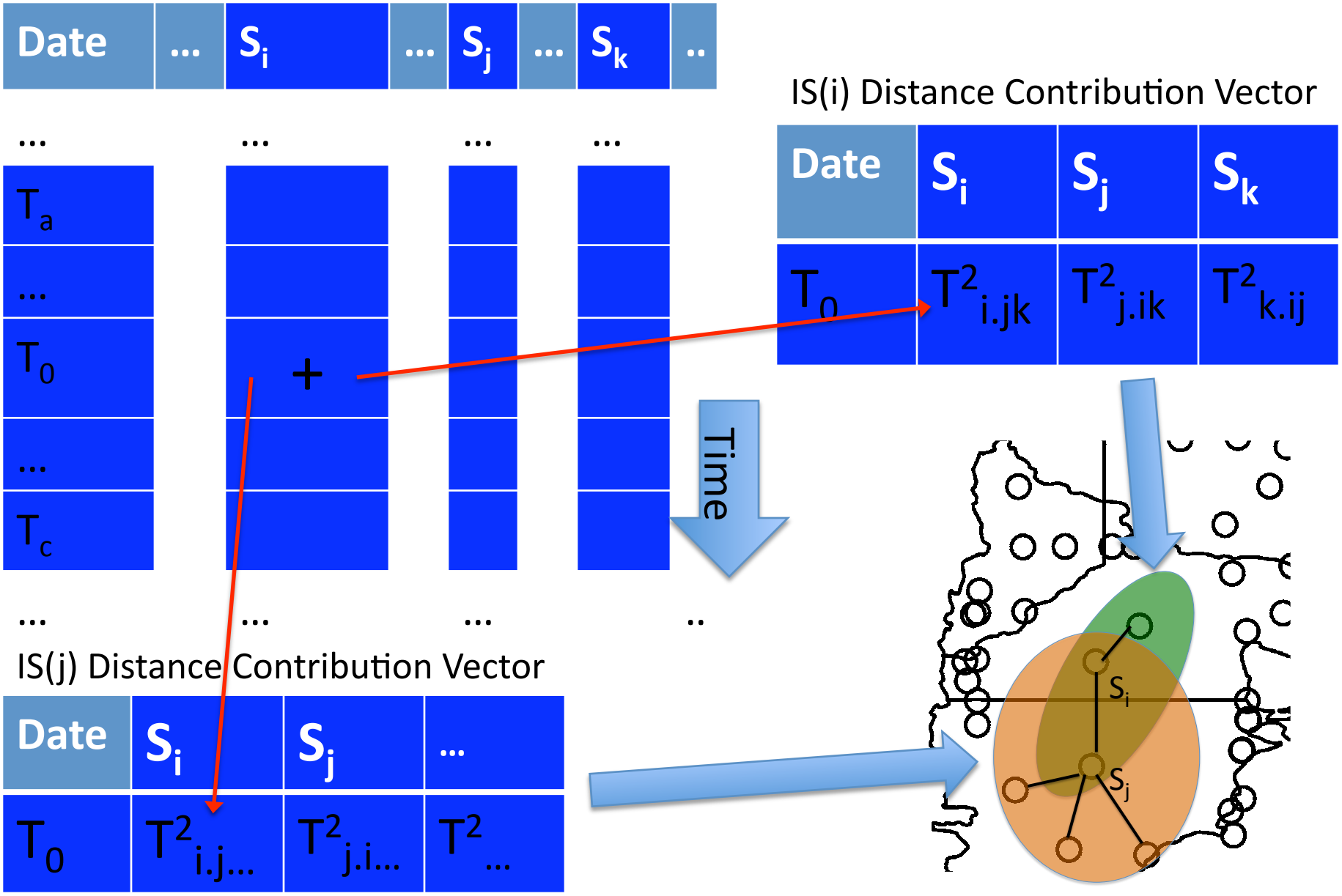
$$IS_o = \{S_o, S_p\}$$
$$IS_p = \{S_p, S_o\}$$
$$IS_i = \{S_i, S_j, S_k\}$$
$$IS_j = \{S_i, S_j, S_l, S_m, S_n\}$$
$$IS_k = \{S_i, S_k\}$$
$$IS_l = \{S_j, S_l\}$$
$$IS_m = \{S_j, S_m\}$$
$$IS_n = \{S_j, S_n\}$$

Both Influence Sets are identical

Two identical distances are relevant here

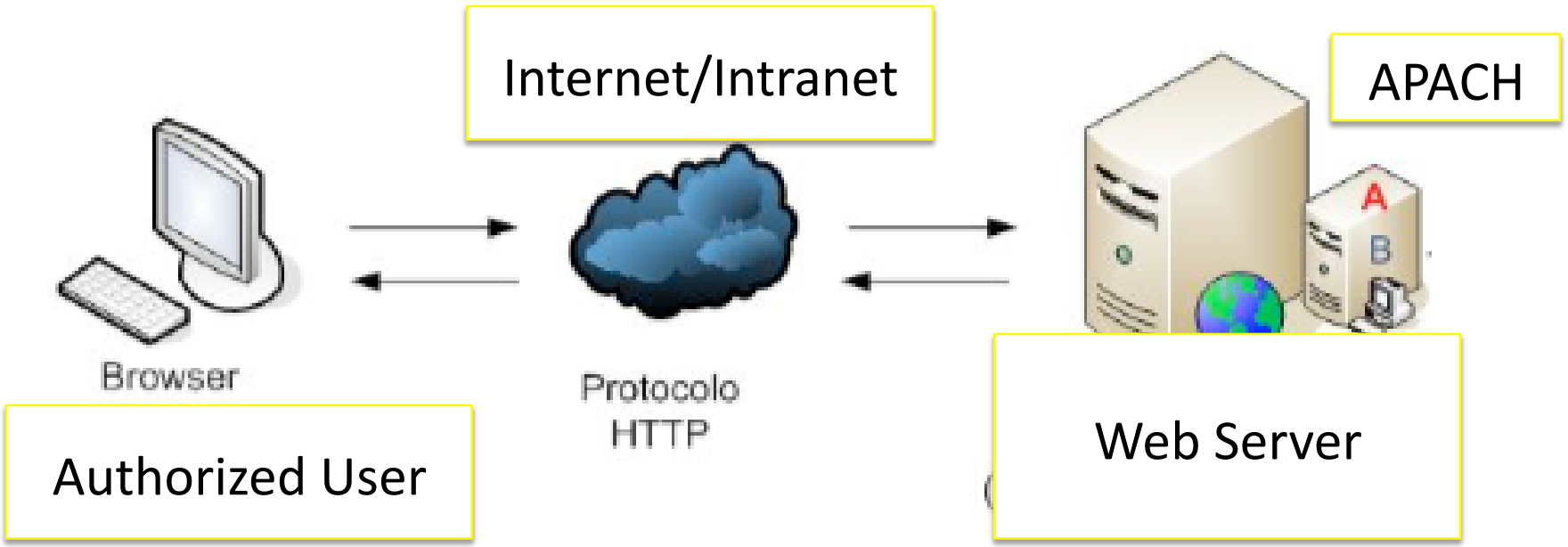
$$T^2_{o/p}$$
$$T^2_{o/p}$$

Distance Contribution Matrix

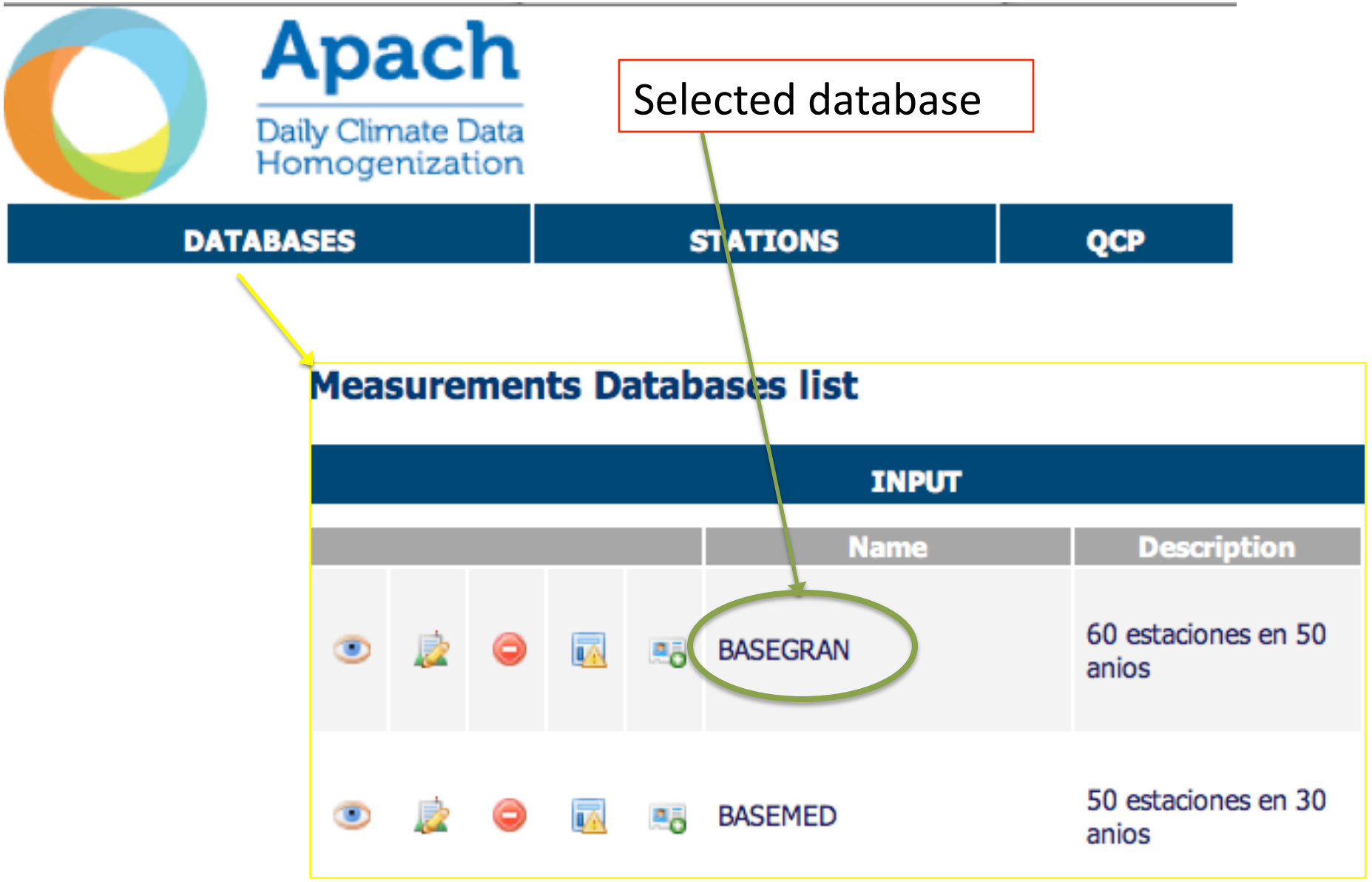


The Algorithms Implementation

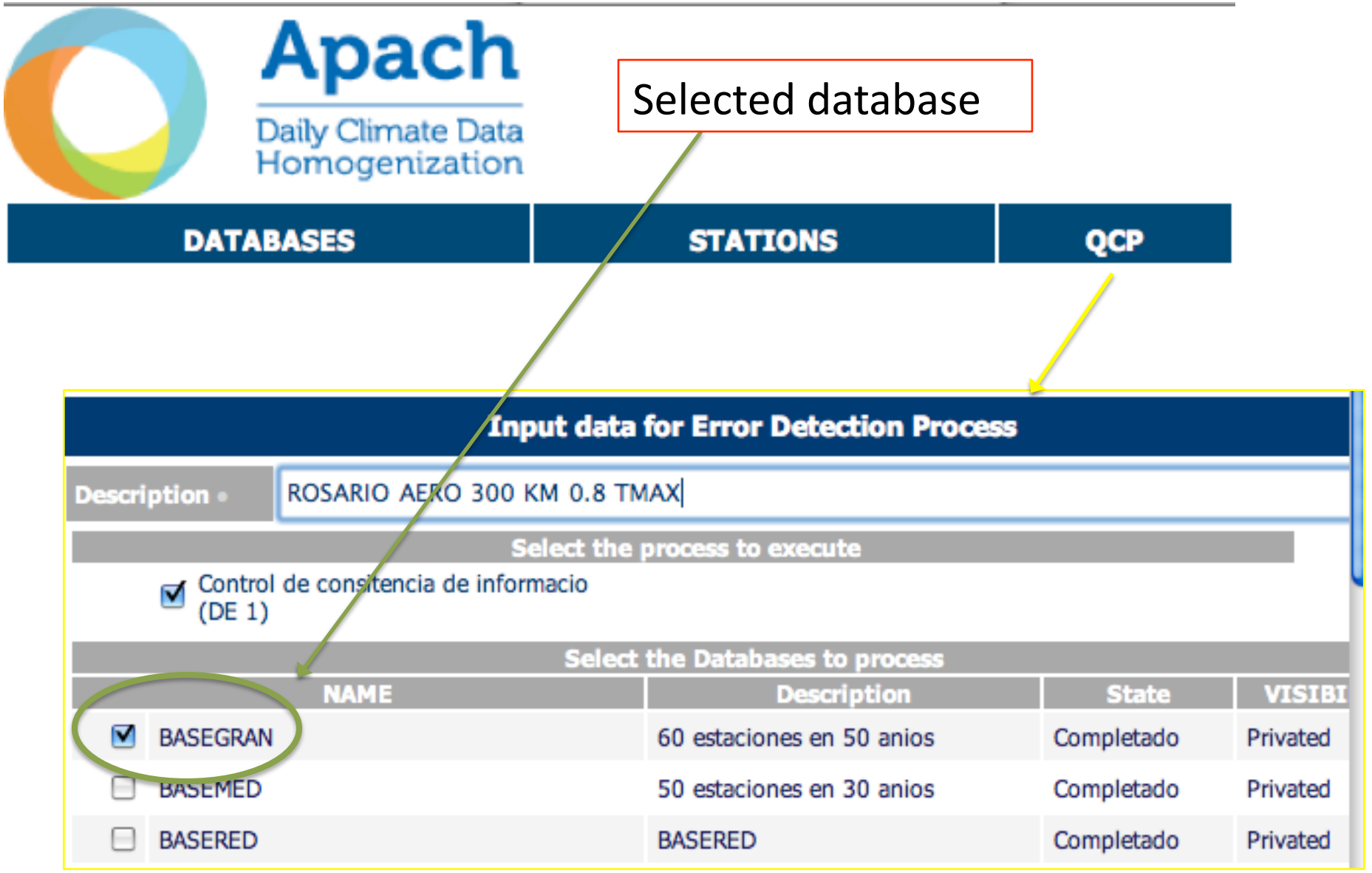
- Algorithms Programmed in R.
- WEB Interface Implemented in PHP.
- Methods Available from the Internet.
- Accesible Wordlwide (Authorized Users).



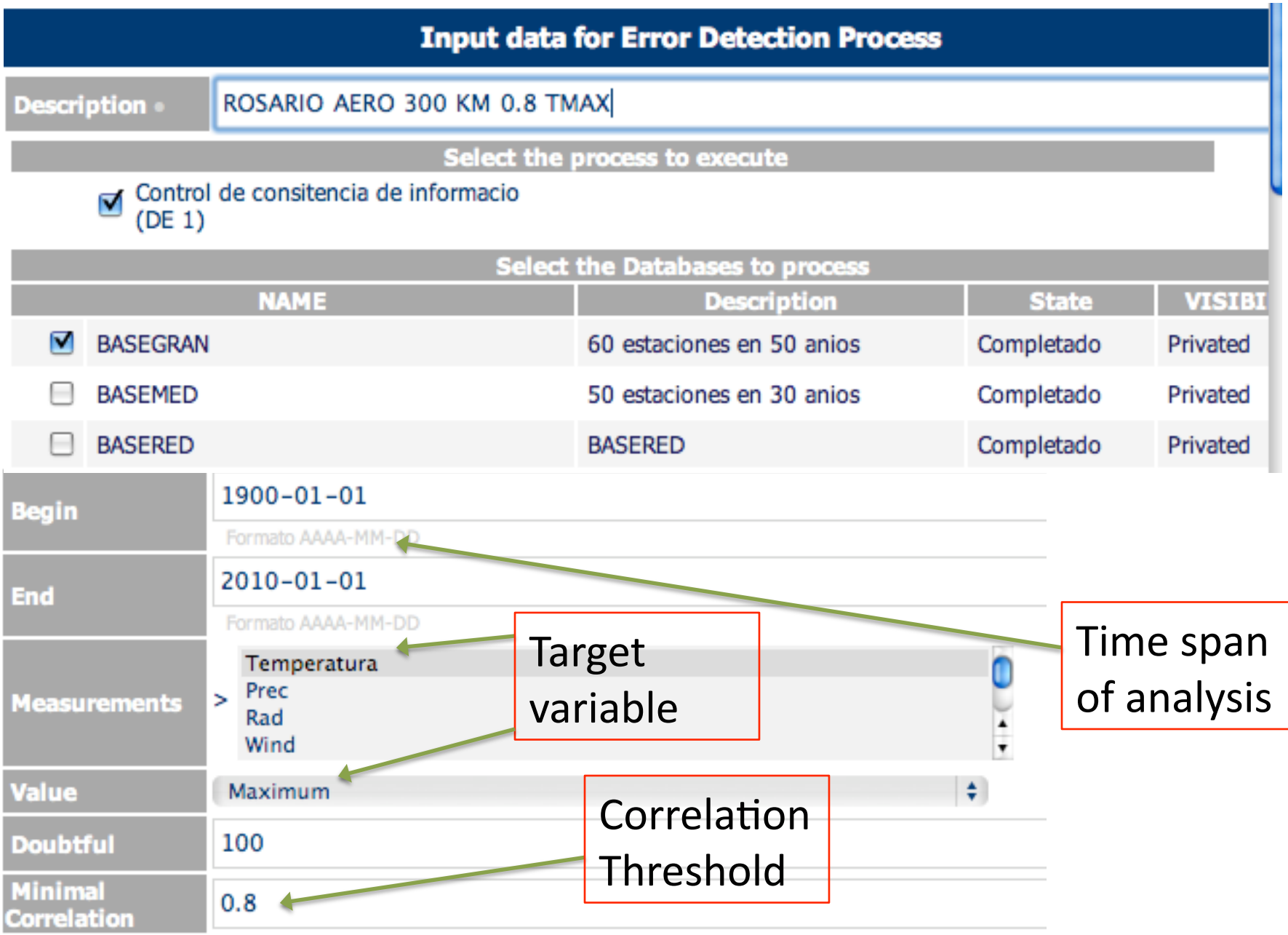
Uploading a database



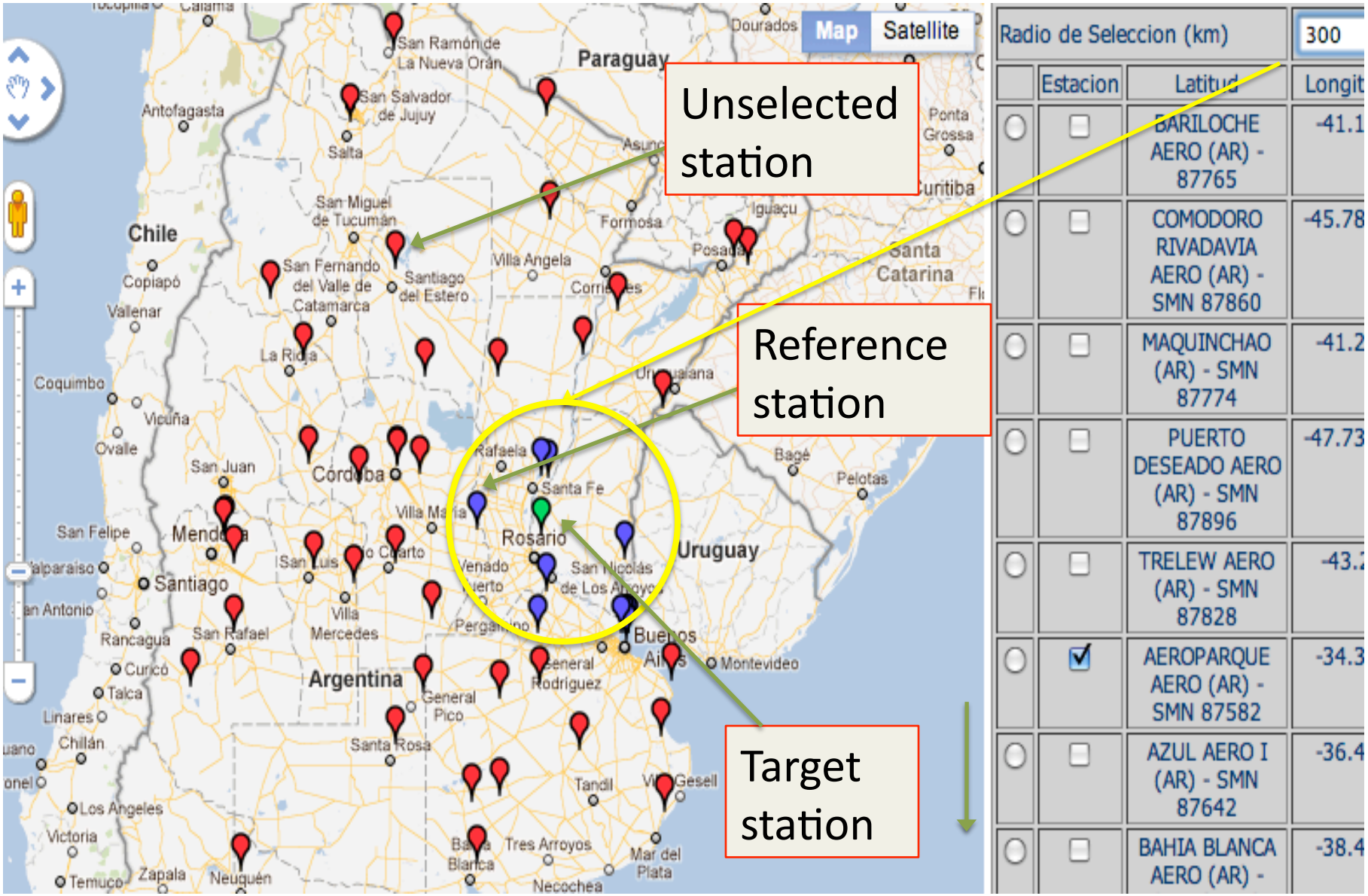
Executing a Quality Control Process



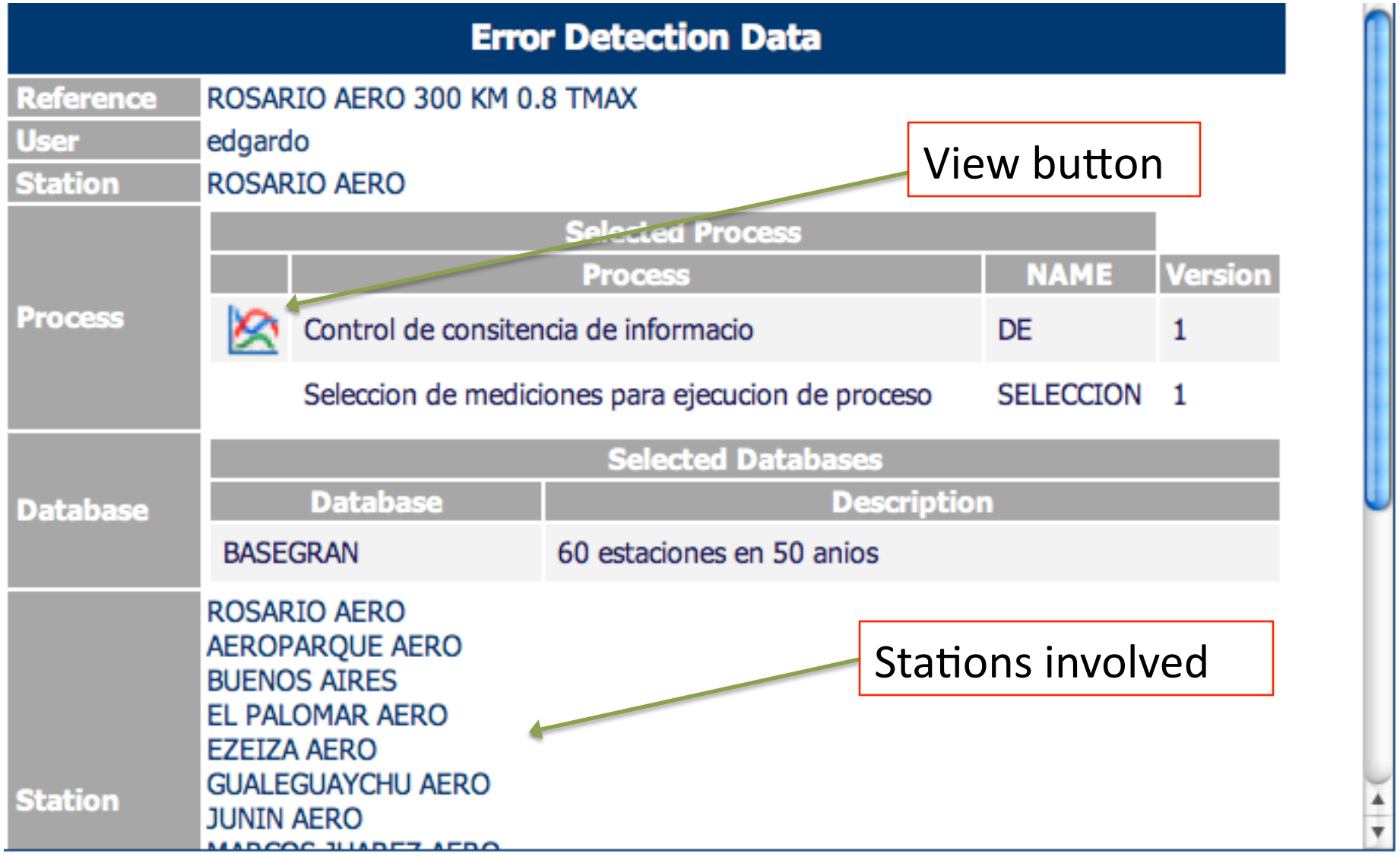
QCP: Parameters determination



Selecting the stations



Executing a Quality Control Process



List of potential errors

Mahalanobis Distance (T^2)

Apach

Daily Climate Data Homogenization

Suspect date

# of stations

Observed Results Measurements

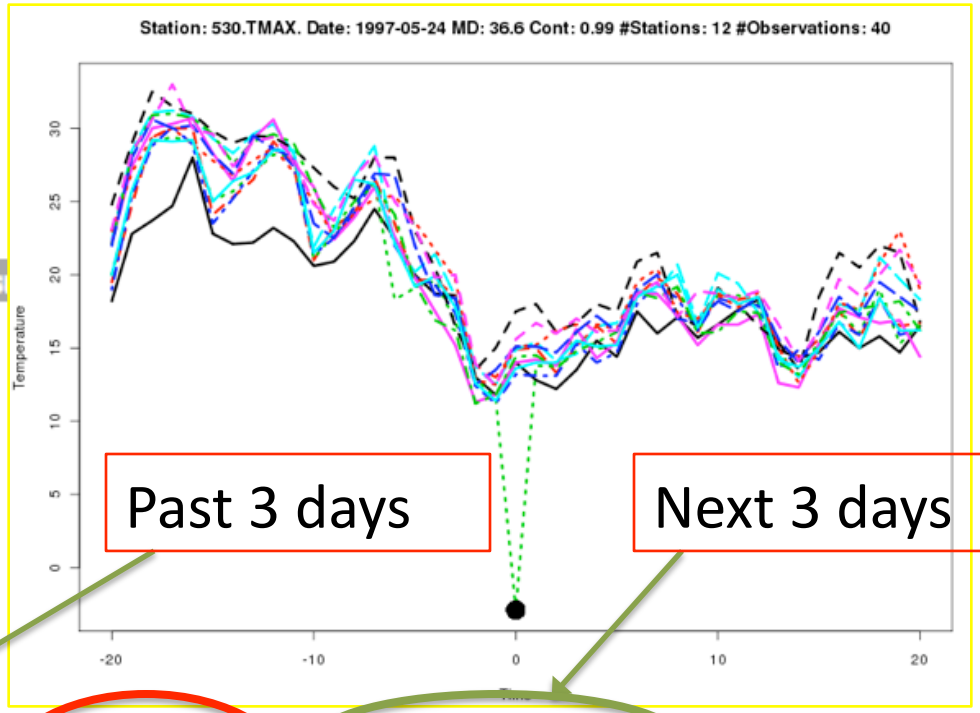
Order by: Severity | Max Contribution

Severity	MD	Max Cont.	Date	Observed	Station
1	44.047	0.991	1955-07-04	41	6
2	36.598	0.992	1997-05-24	40	12
3	30.545	0.943	1974-05-25	41	12
4	30.264	0.921	1956-02-19	29	8
5	27.809	0.938	1953-10-31	35	5
6	26.895	0.926	1984-11-04	41	12

Ordering buttons

Maximum Contribution

Is this an error ?

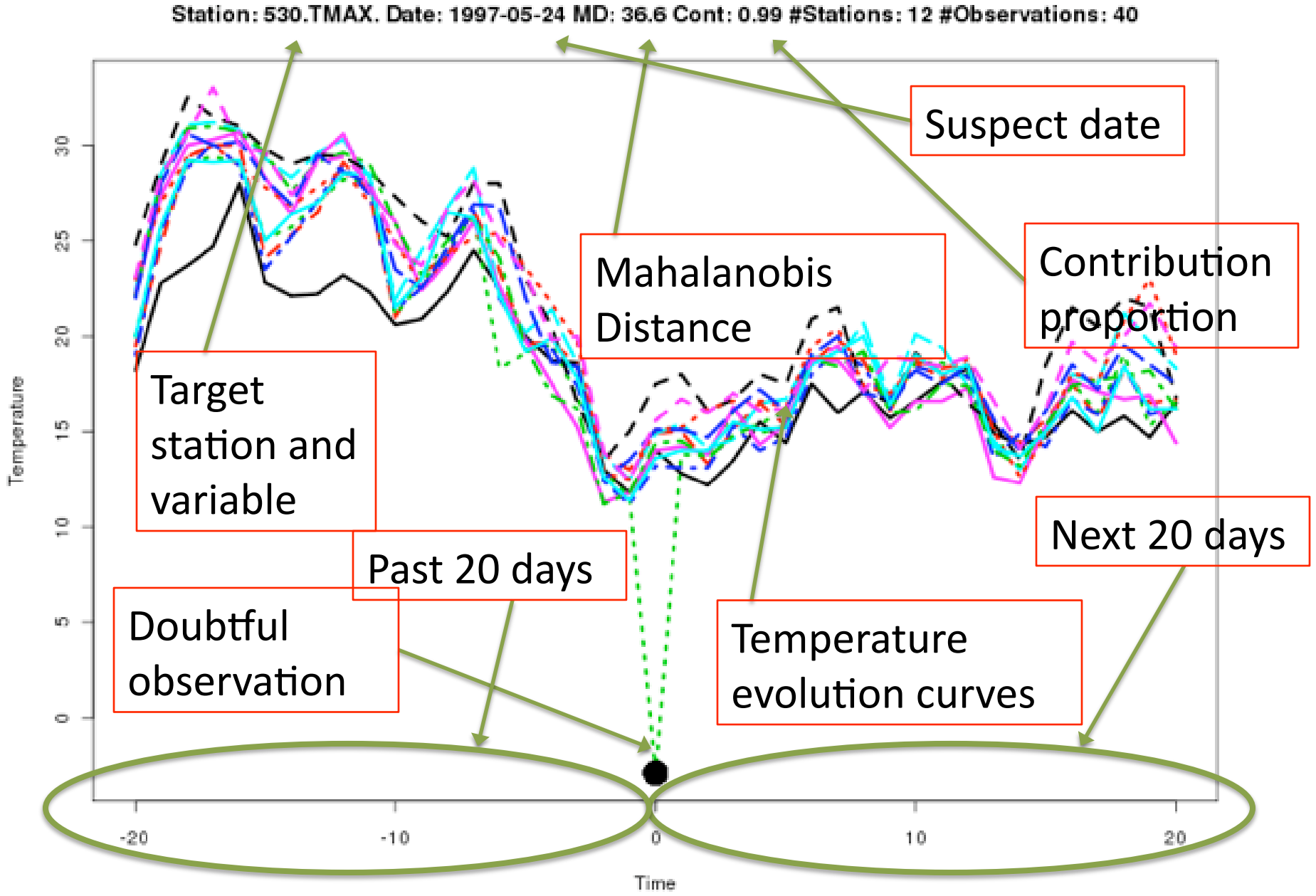


Observed Results Meas	
Severity	2
Date	1997-05-24
MD	36.598
Observed	40
Station	12

Station	Cont		-3	-2	-1	1997-05-24		+1	+2	+3
EL PALOMAR AERO (-34.36 , -58.36 , 1)	0.992	Min Max	14.8 16.800	4.8 12.900	0.8 11.600	13.9 -2.900	Editar	3.5 13.800	-0.3 13.700	3.2 16.100
SAN MIGUEL (-34.33 , -58.44 , 1)	0.183	Min Max	12 17.800	6 12.600	1.5 11.400	0.4 13.600	Editar	3.6 14.000	2.8 14.000	6.6 15.500
EZEIZA AERO (-34.49 , -58.32 , 1)	0.064	Min Max	12 17.400	5.1 12.400	0.2 11.200	-1.6 13.200	Editar	2.8 13.100	1.2 13.100	5 15.300
BUENOS AIRES (-34.35 , -58.29 , 1)	0.049	Min Max	11.7 18.500	6.2 13.800	3.2 13.000	1.4 14.900	Editar	4.7 15.000	4.2 13.300	8 15.100
GUALEGUAYCHU AERO (-33 , -58.37 , 1)	0.007	Min Max	12.5 18.500	6.5 13.700	0.9 12.500	3.7 14.900	Editar	2.5 15.400	3.9 14.100	5.6 14.600
MARCOS JUAREZ AERO (-32.42 , -62.09 , 1)	0.004	Min Max	9.5 16.500	1.7 13.500	-2.6 15.000	1 17.500	Editar	-0.7 18.000	2.2 16.100	5.3 16.700
JUNIN AERO (-34.33 , -60.55 , 1)	0.002	Min Max	10.6 15.200	3.2 11.300	-0.7 11.800	-0.4 14.000	Editar	3.1 14.200	3.6 13.800	4 16.300

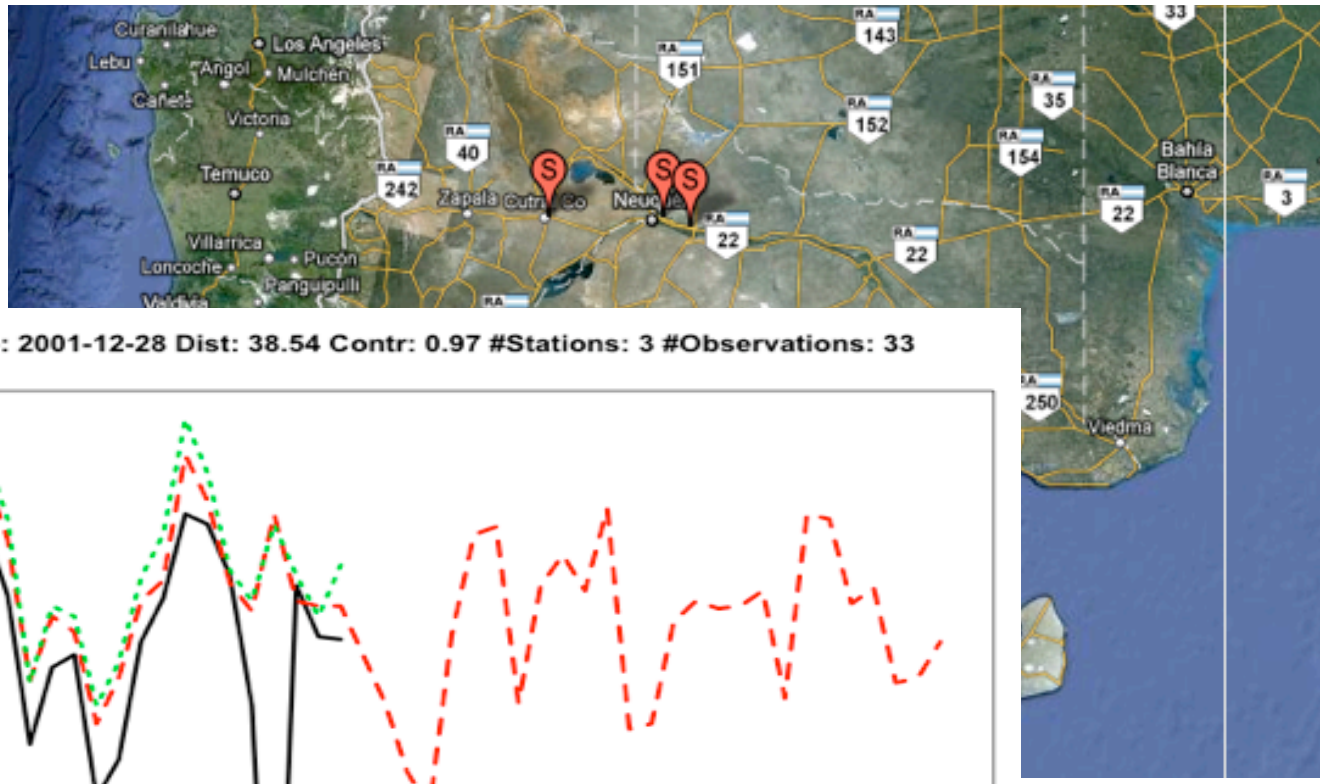
Tmin and Tmax mixed up

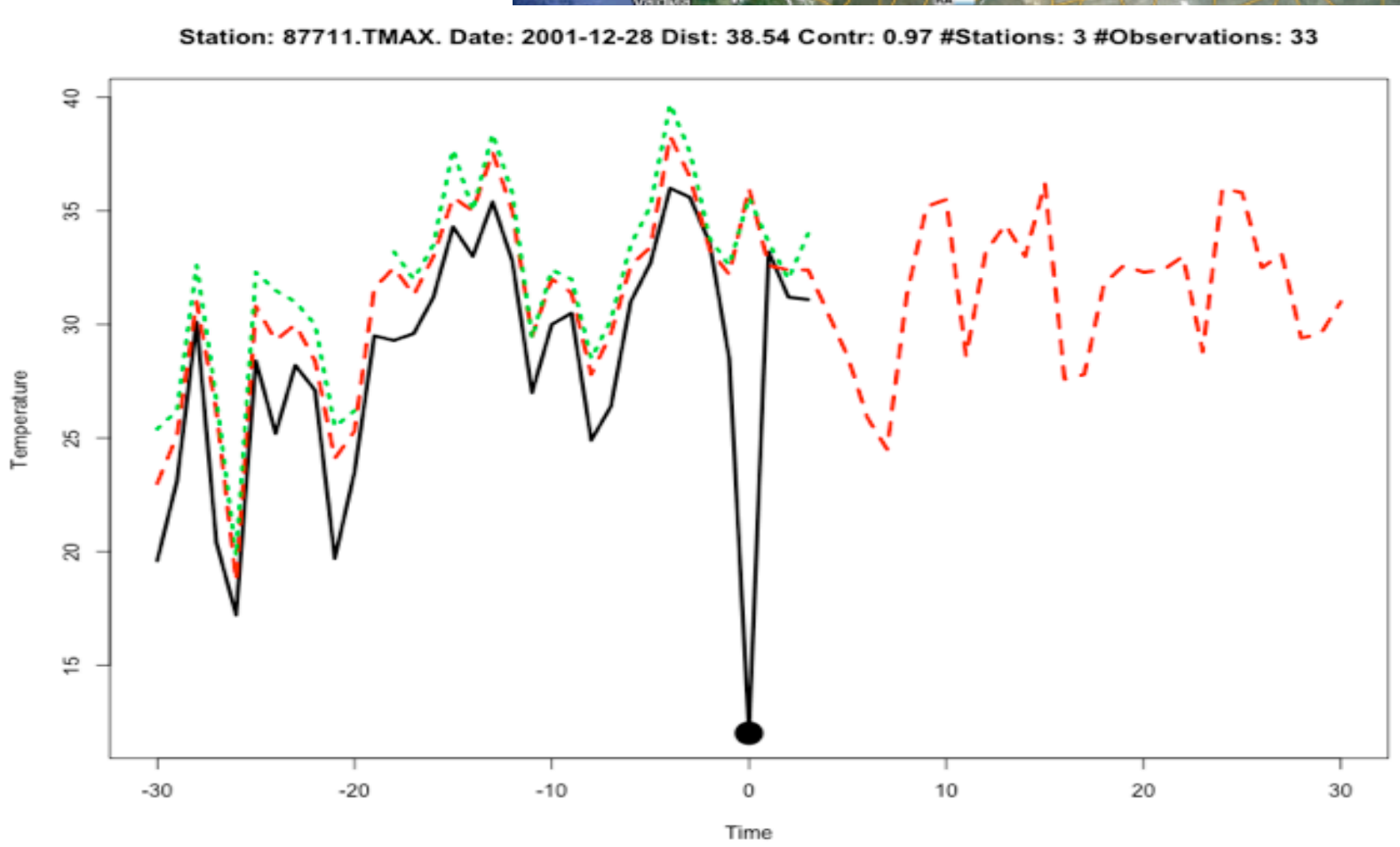
The Graphical Output in detail



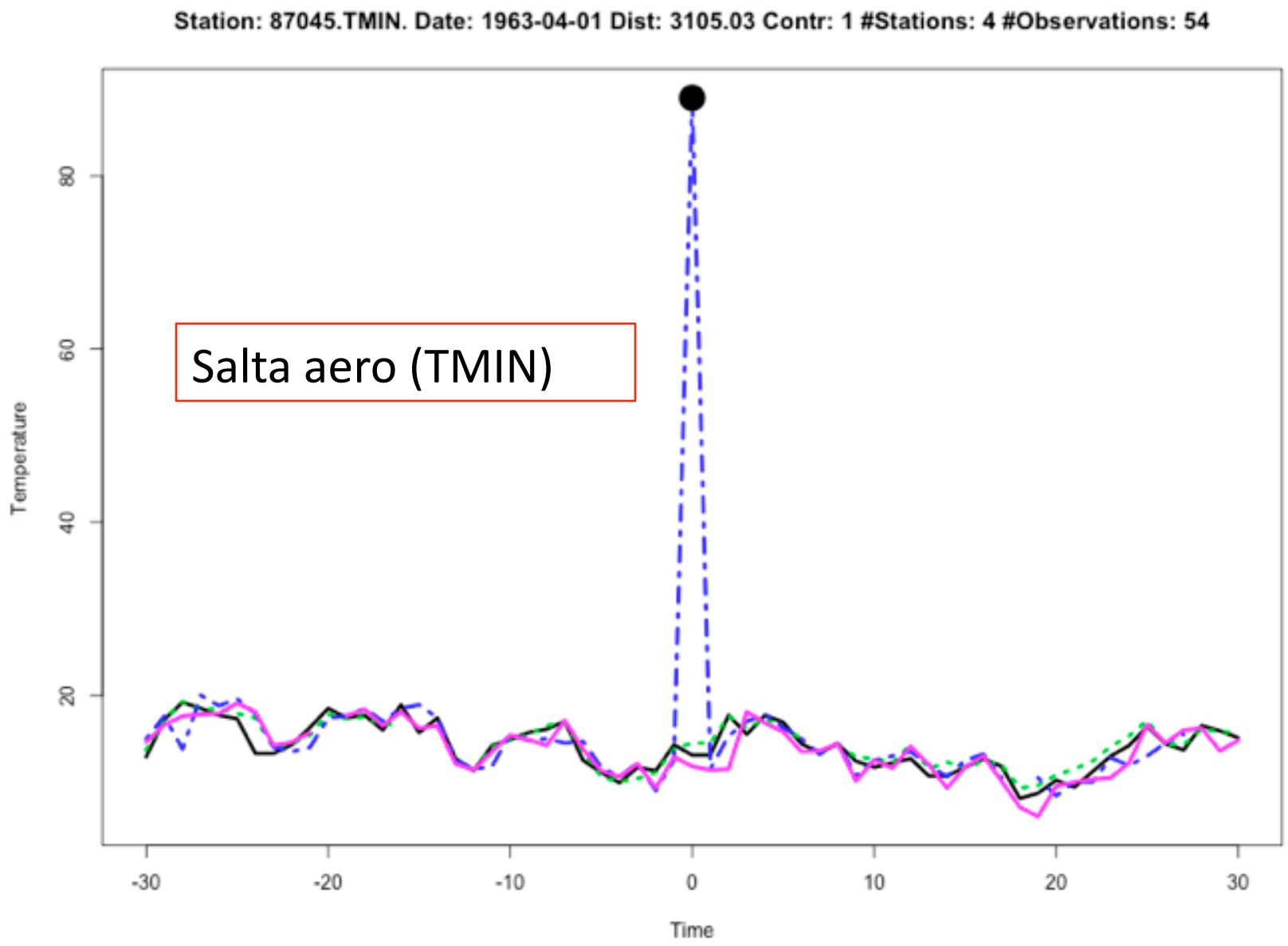
Some Examples

Univariate Outlier (3 Stations)

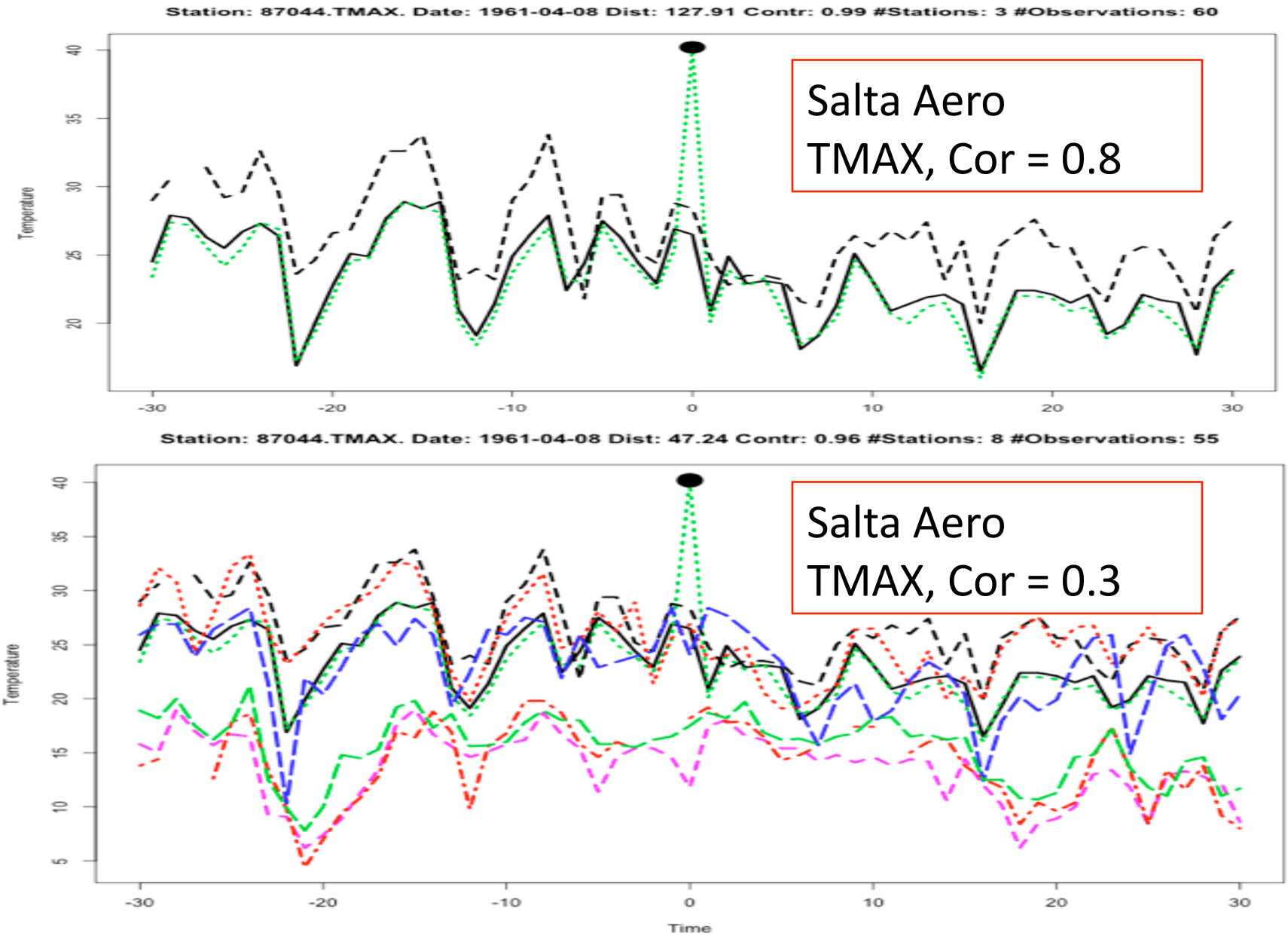




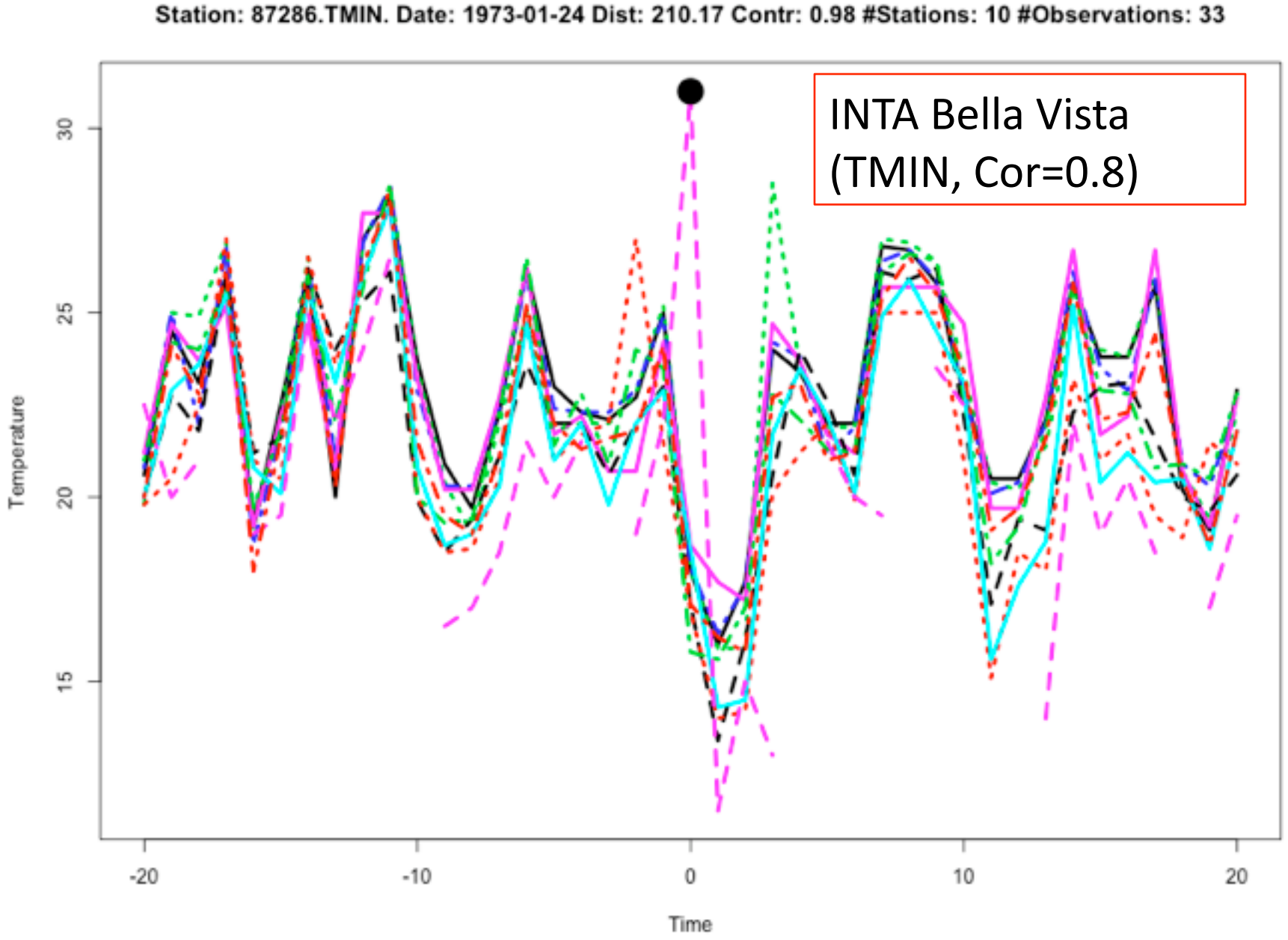
Huge Univariate Outlier



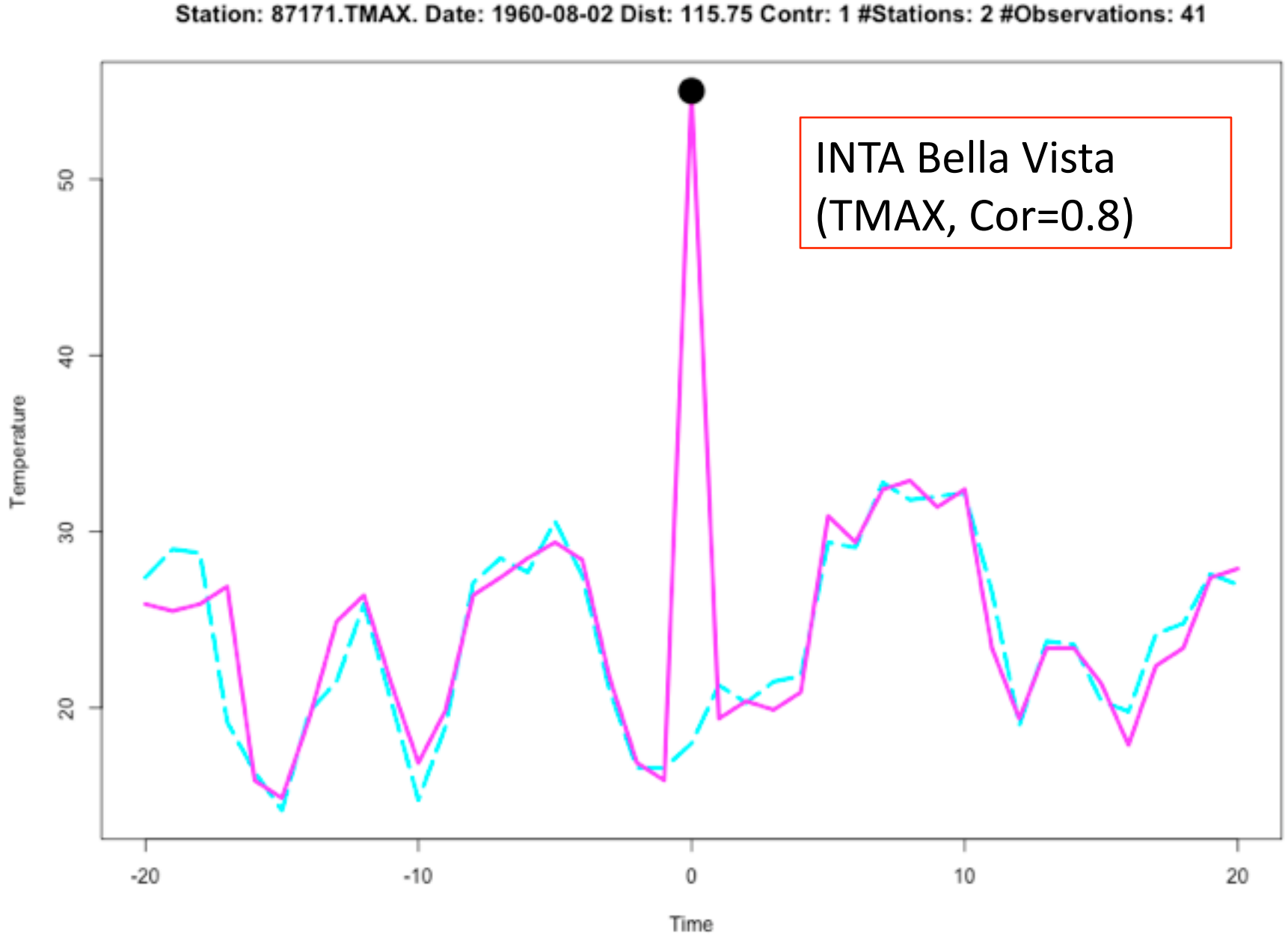
Changing the Correlation



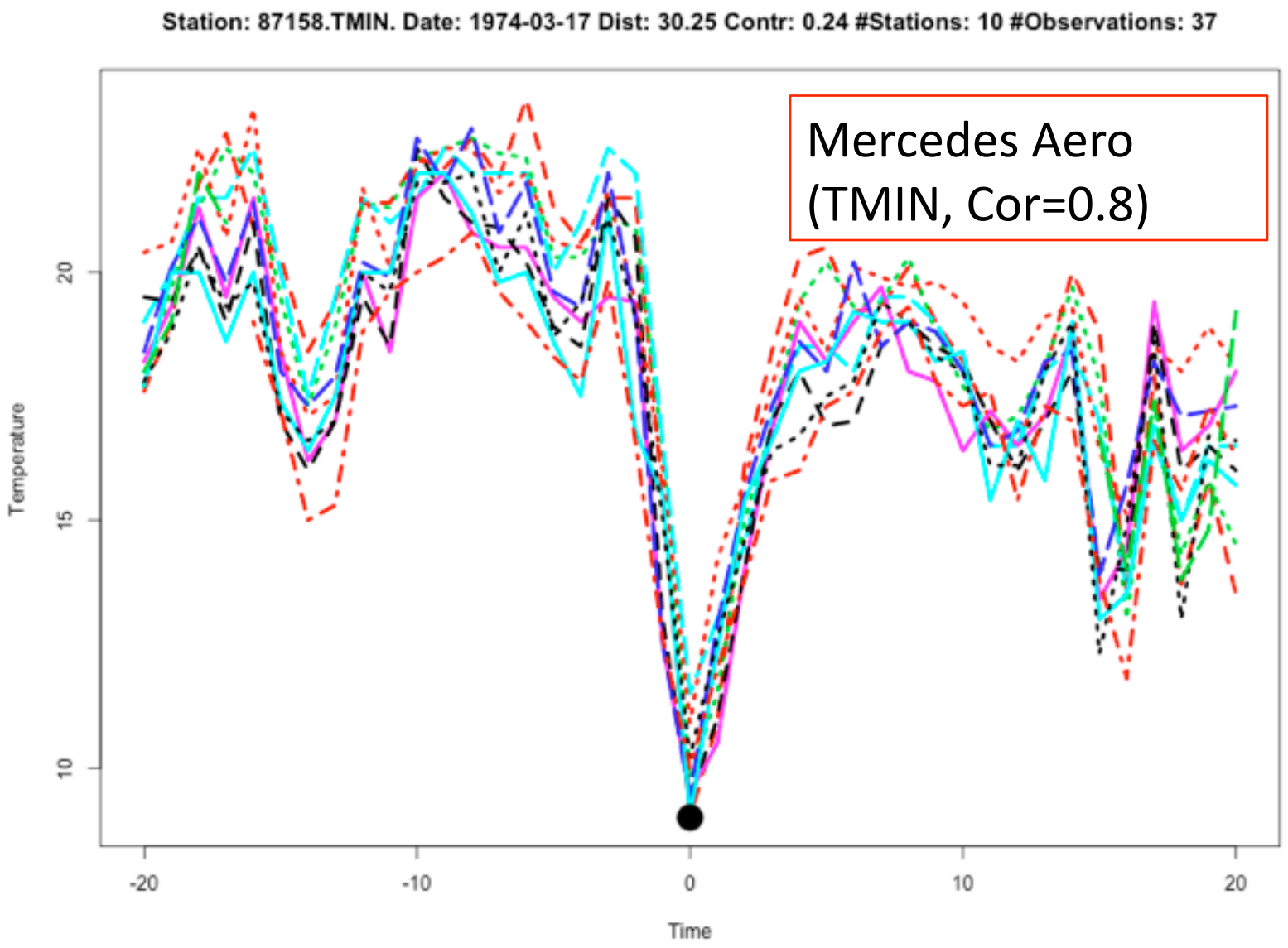
Potential Delay in Temperature Change



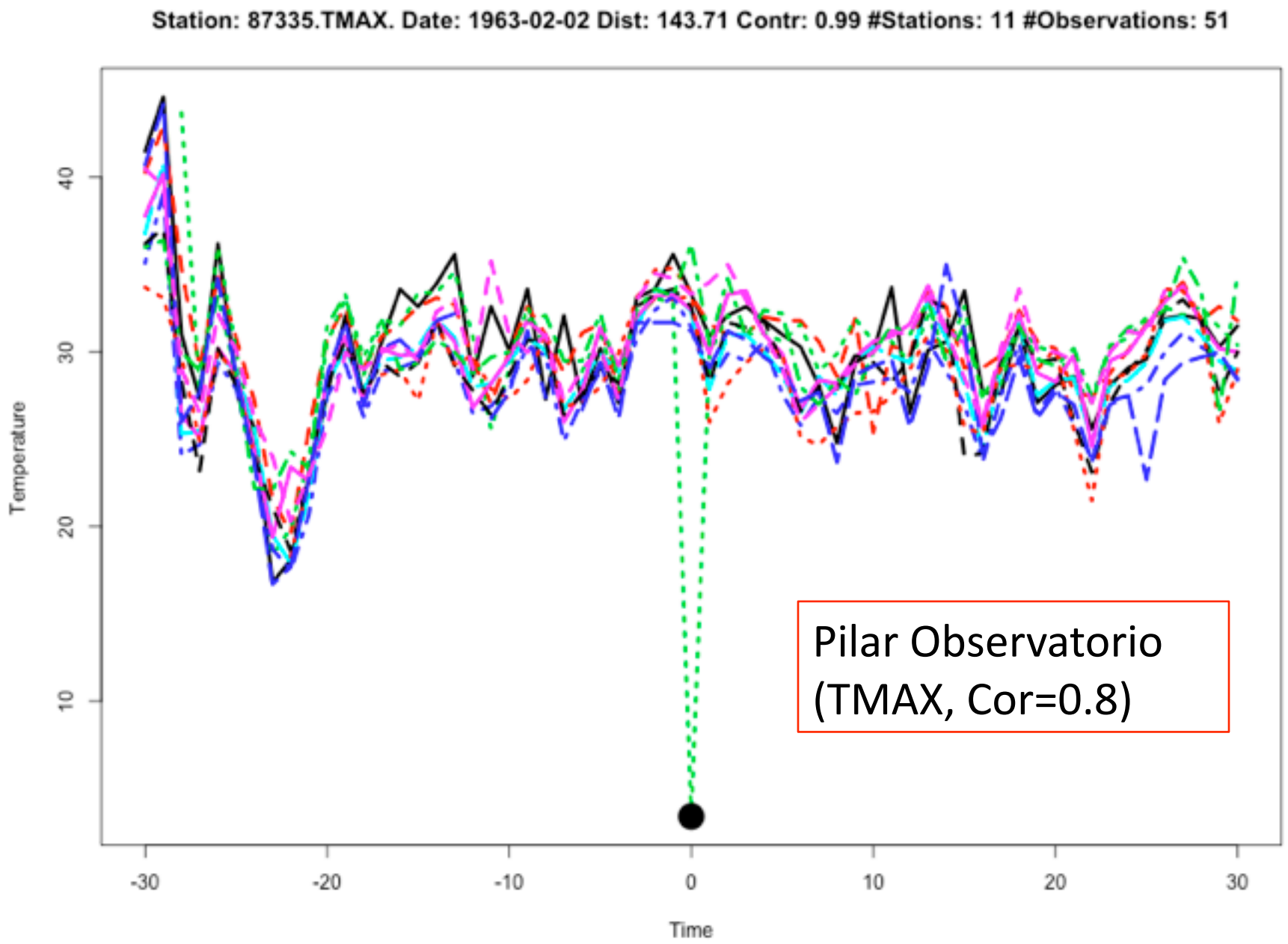
Huge Univariate Outlier



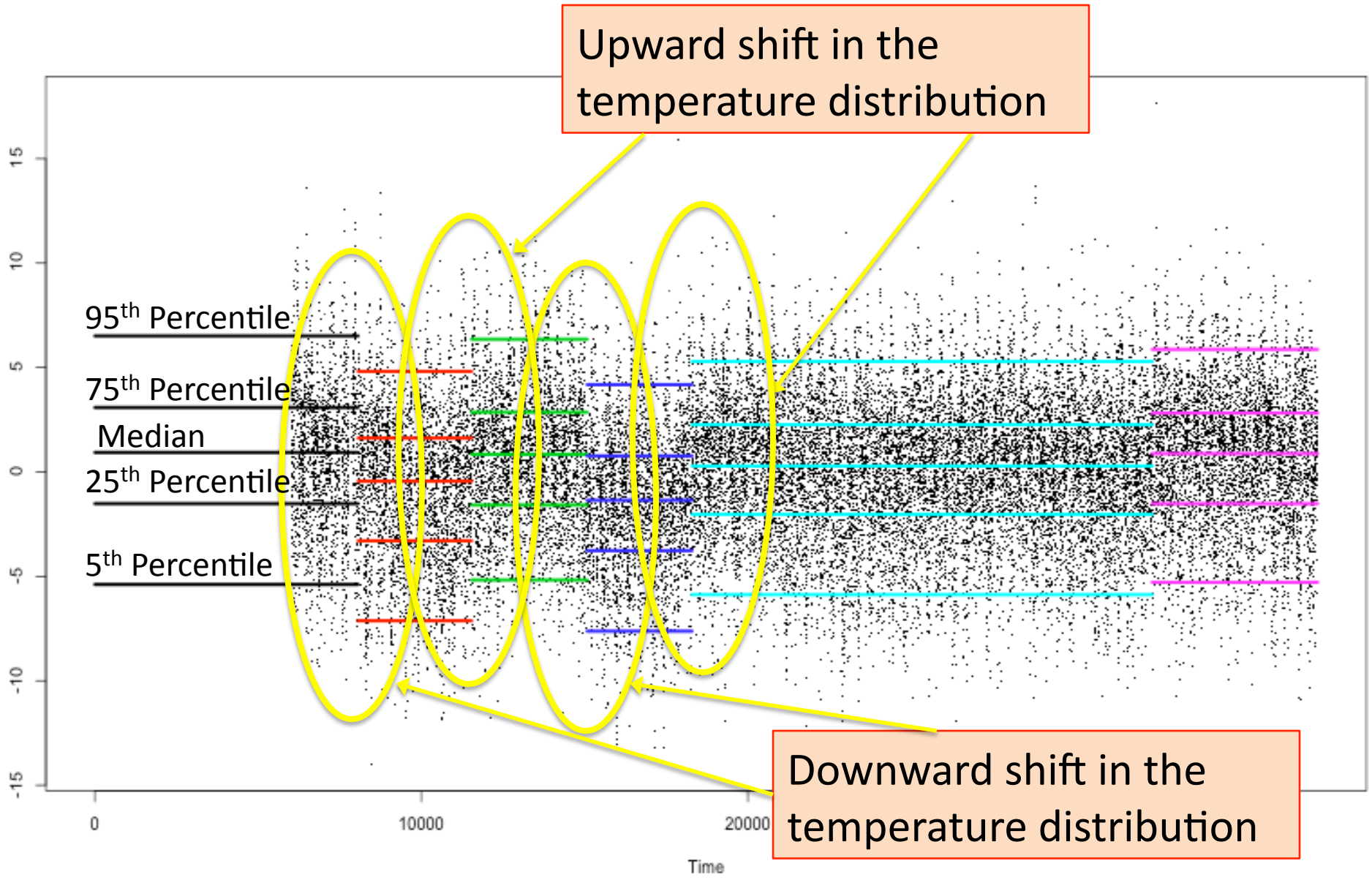
Multivariate Outlier: Odd Vs. Wrong



Zero Instead of Missing Value



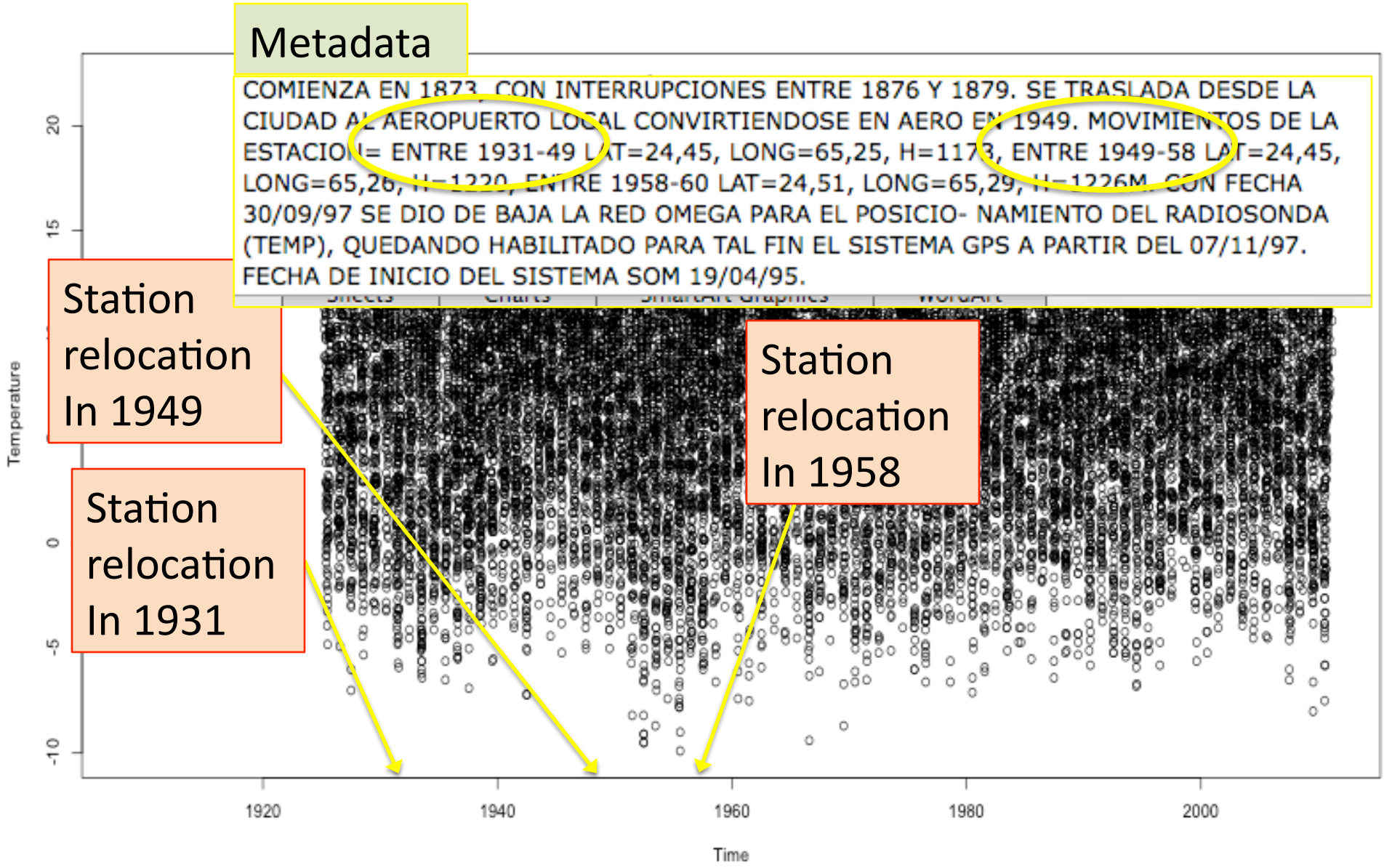
What is an Inhomogeneity



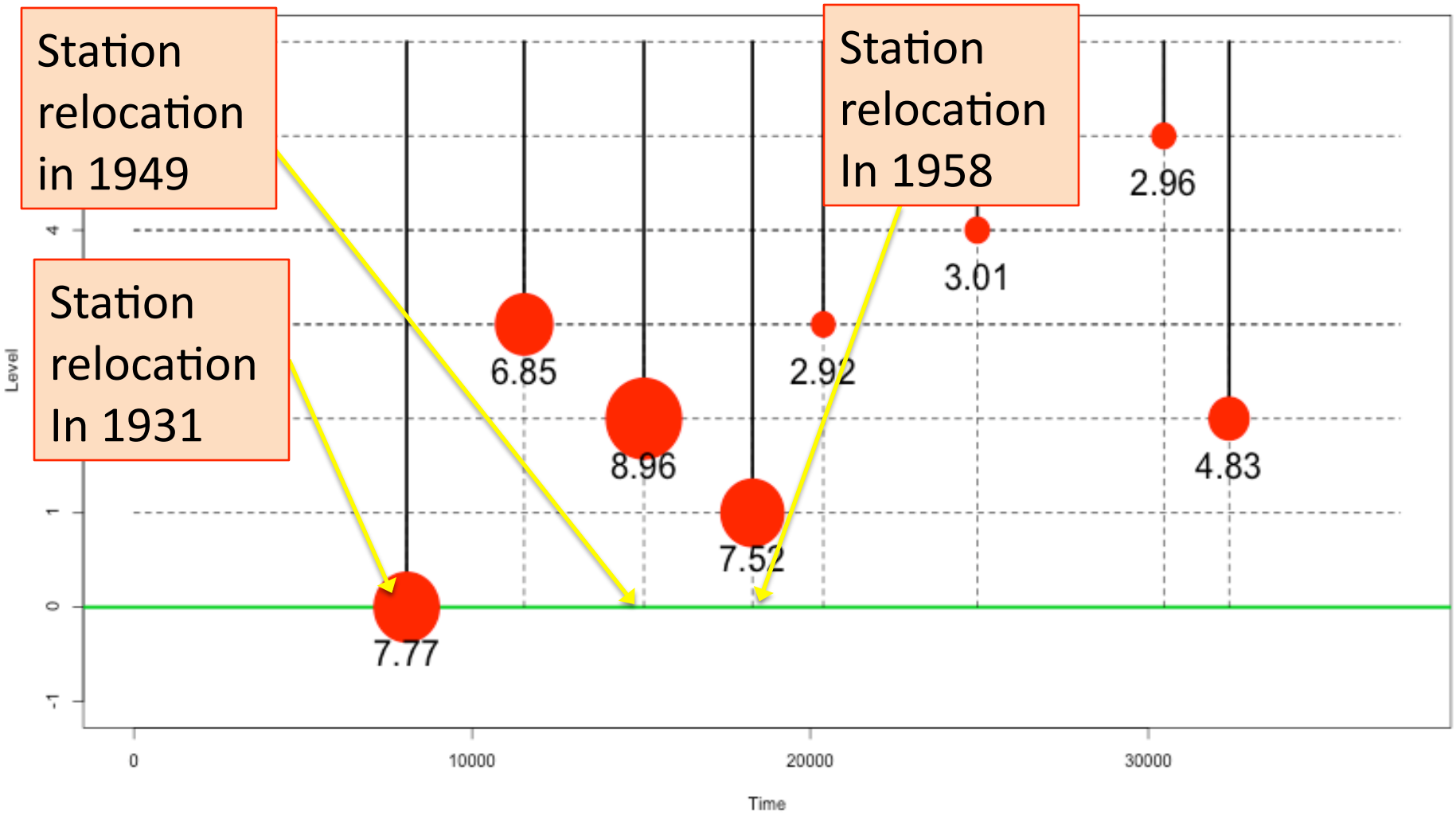
Inhomogeneties Detection Method

- Still under development
- Capable of:
  - Working with multiple stations simultaneausly
  - Identifying the culprit station/s
  - Detecting more than one change/inhomogeneity per station
  - Identifying the kind of change/inhomogeneity (mean, variance, etc.)
  - Evaluating the significance of the change/ inhomogeneity

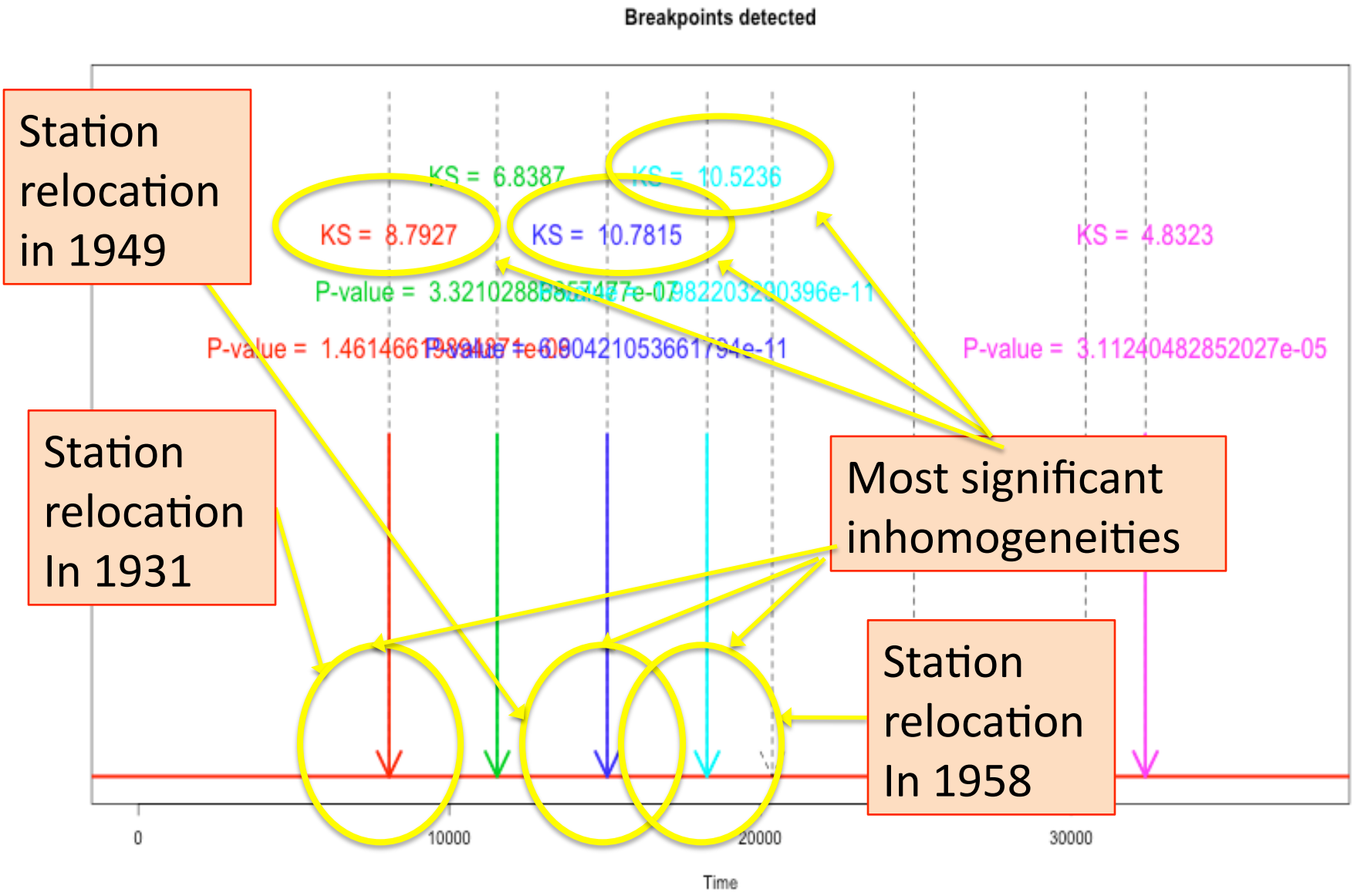
Case study with metadata : Salta Aero (87047, Tmin)



Growing the time partition tree



Significant breakpoints detection



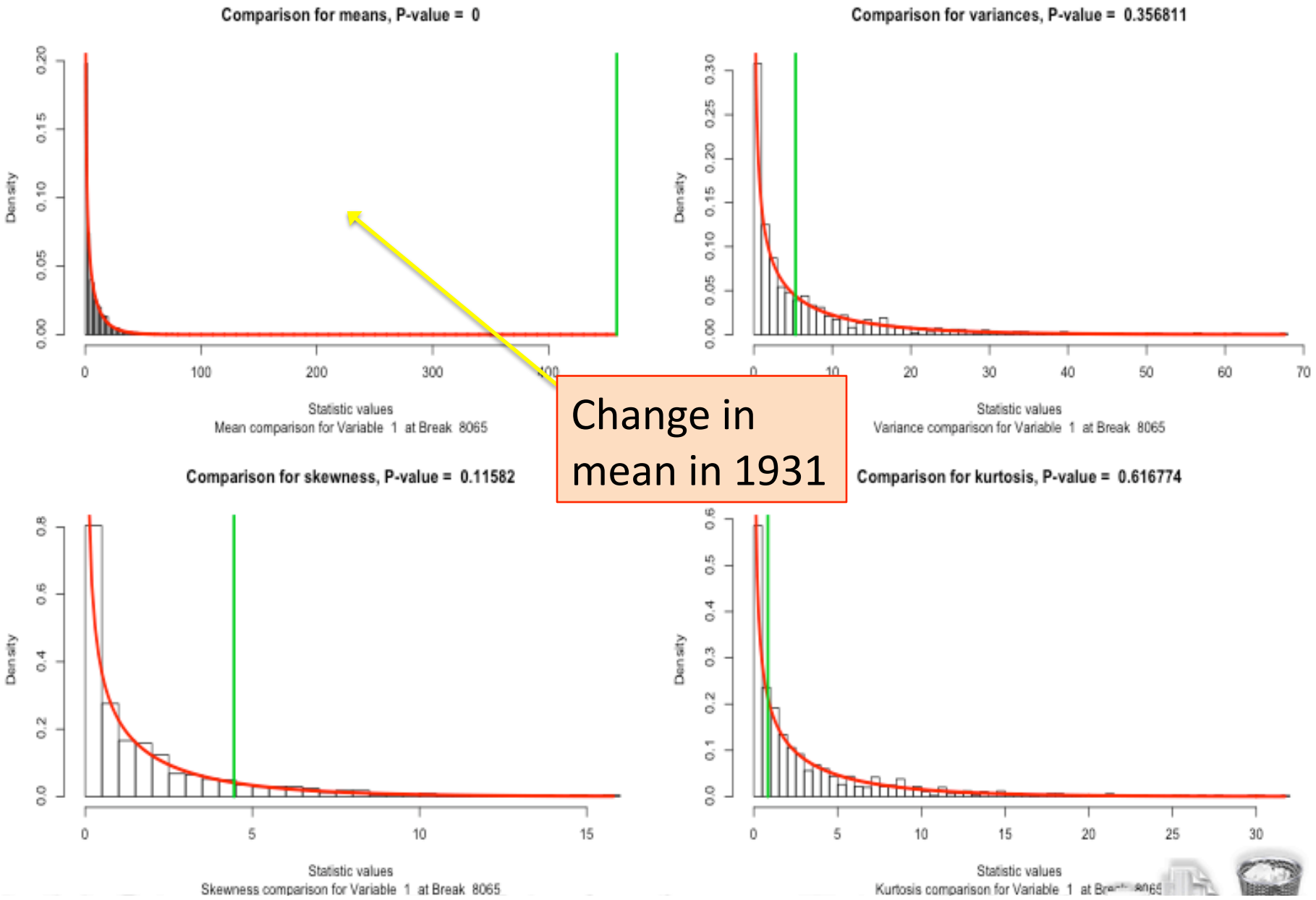
Station relocation in 1949

Station relocation In 1931

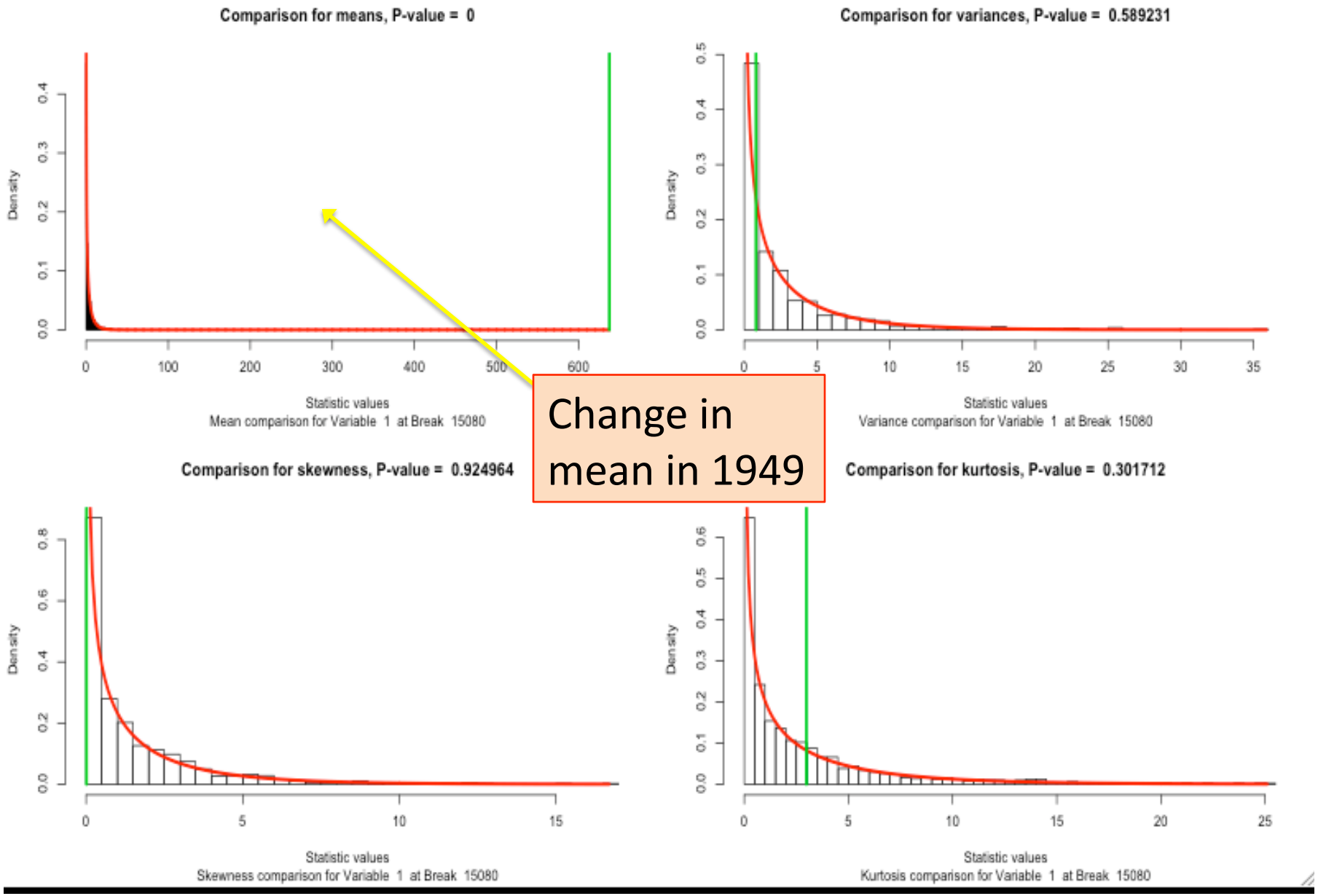
Most significant inhomogeneities

Station relocation In 1958

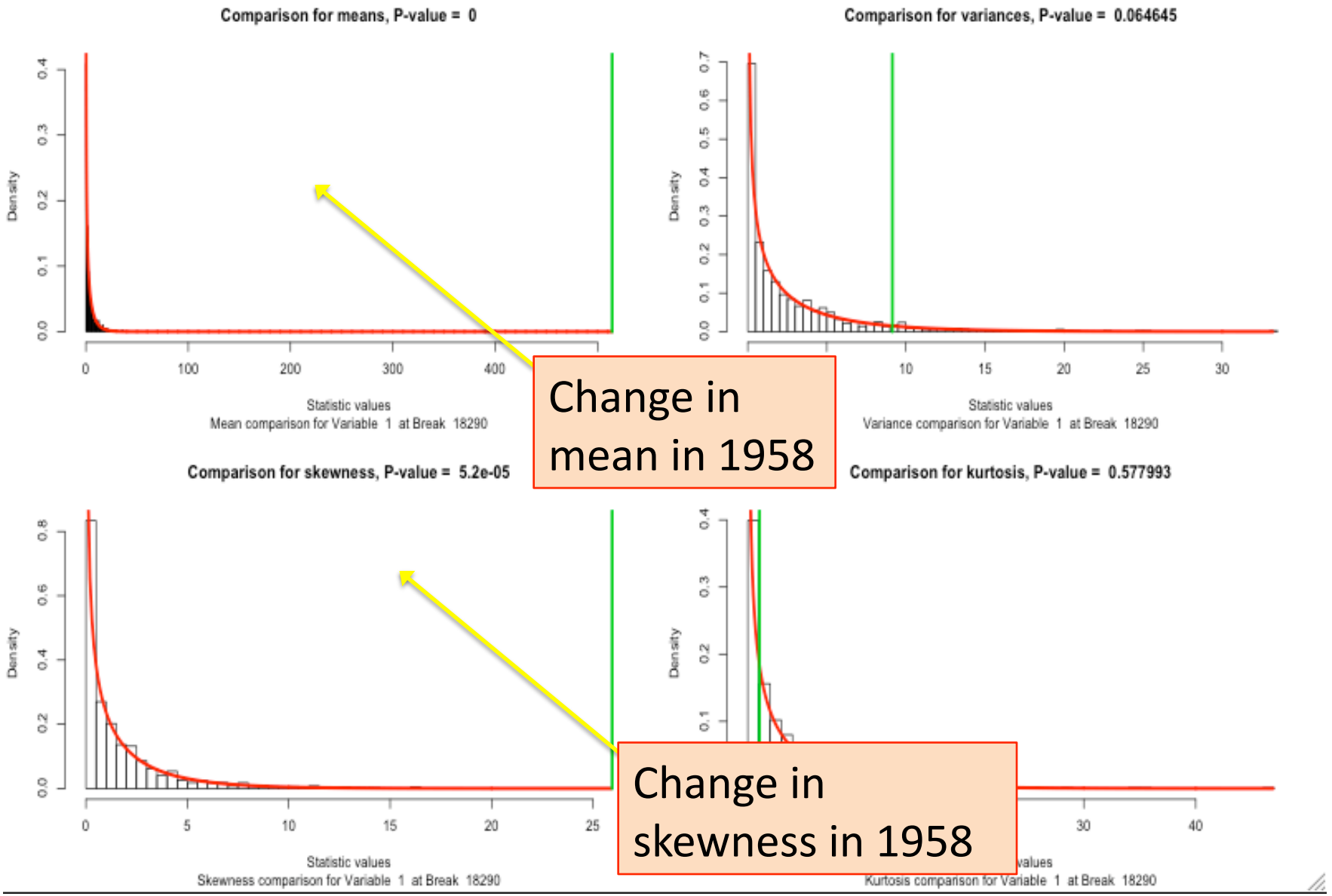
Determining the kind of inhomogeneity



Determining the kind of inhomogeneity



Determining the kind of inhomogeneity

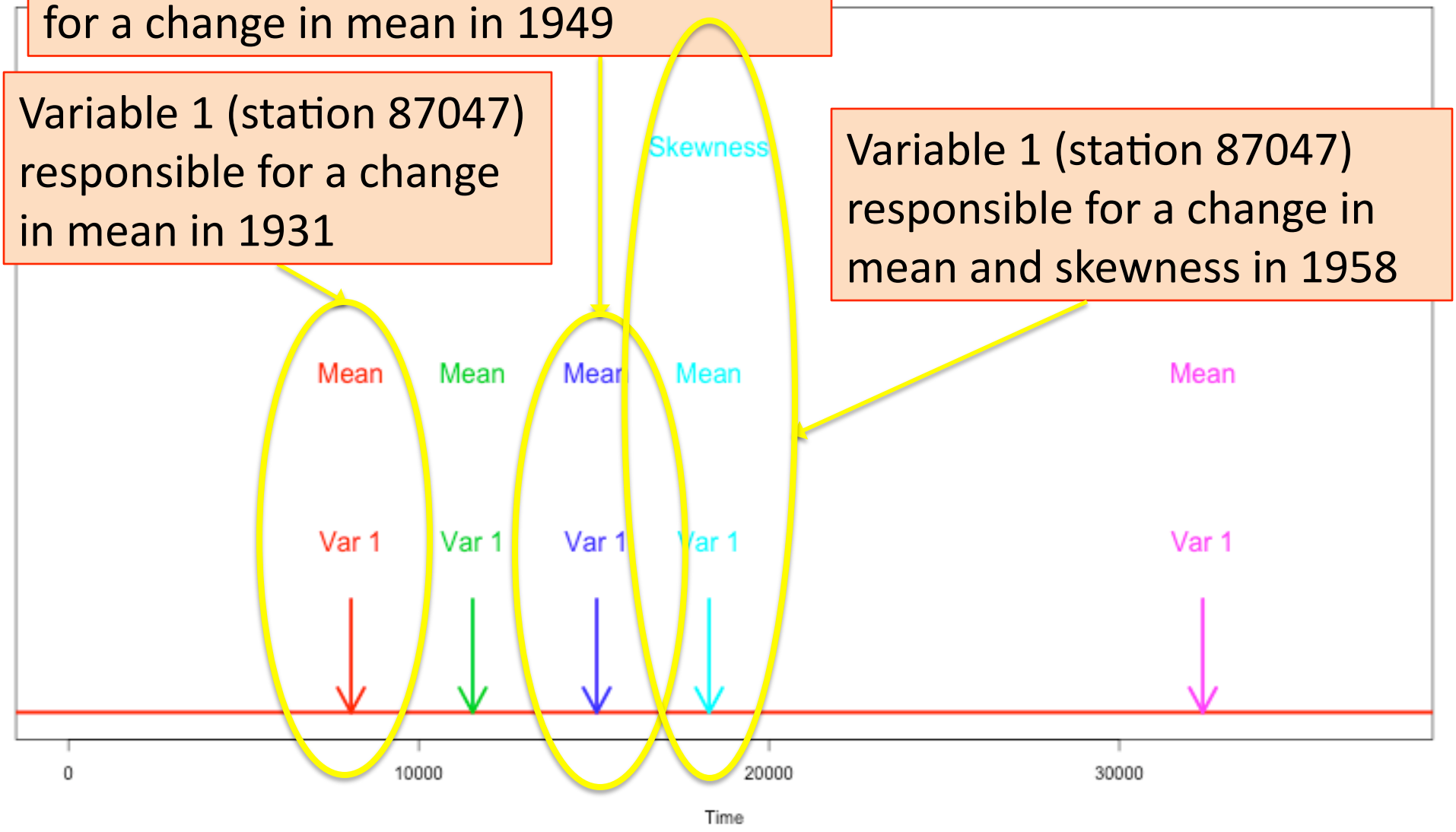


Identification of culprit station and kind of inhomogeneity

Variable 1 (station 87047) responsible for a change in mean in 1949

Variable 1 (station 87047) responsible for a change in mean in 1931

Variable 1 (station 87047) responsible for a change in mean and skewness in 1958



Showing the distributional changes