

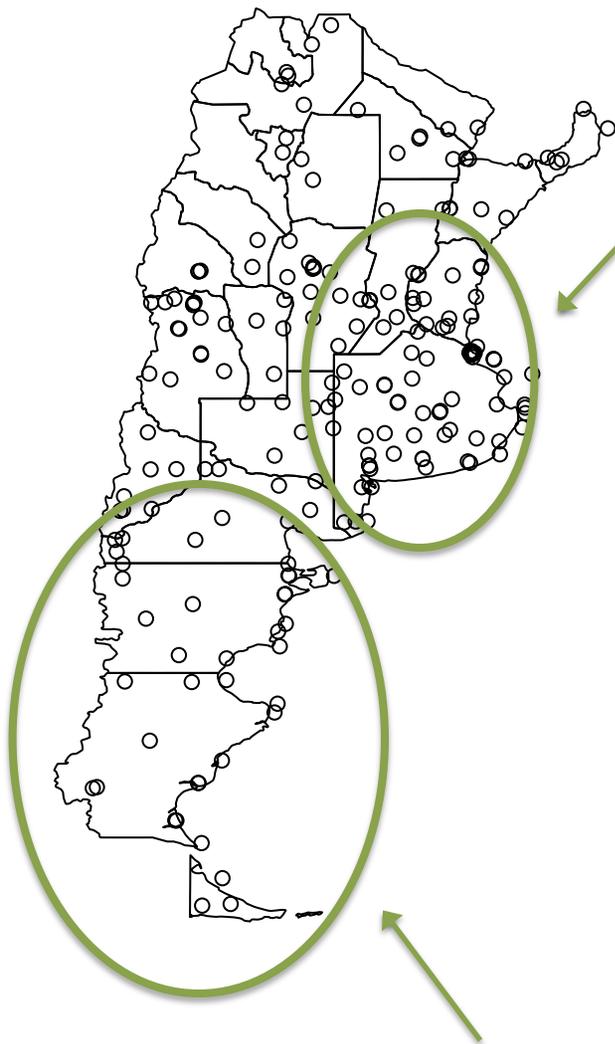
Error and Inhomogeneity detection in daily climate time series

Andrés Farall
Jean-philippe Boulanger
Liliana Orellana

Goals

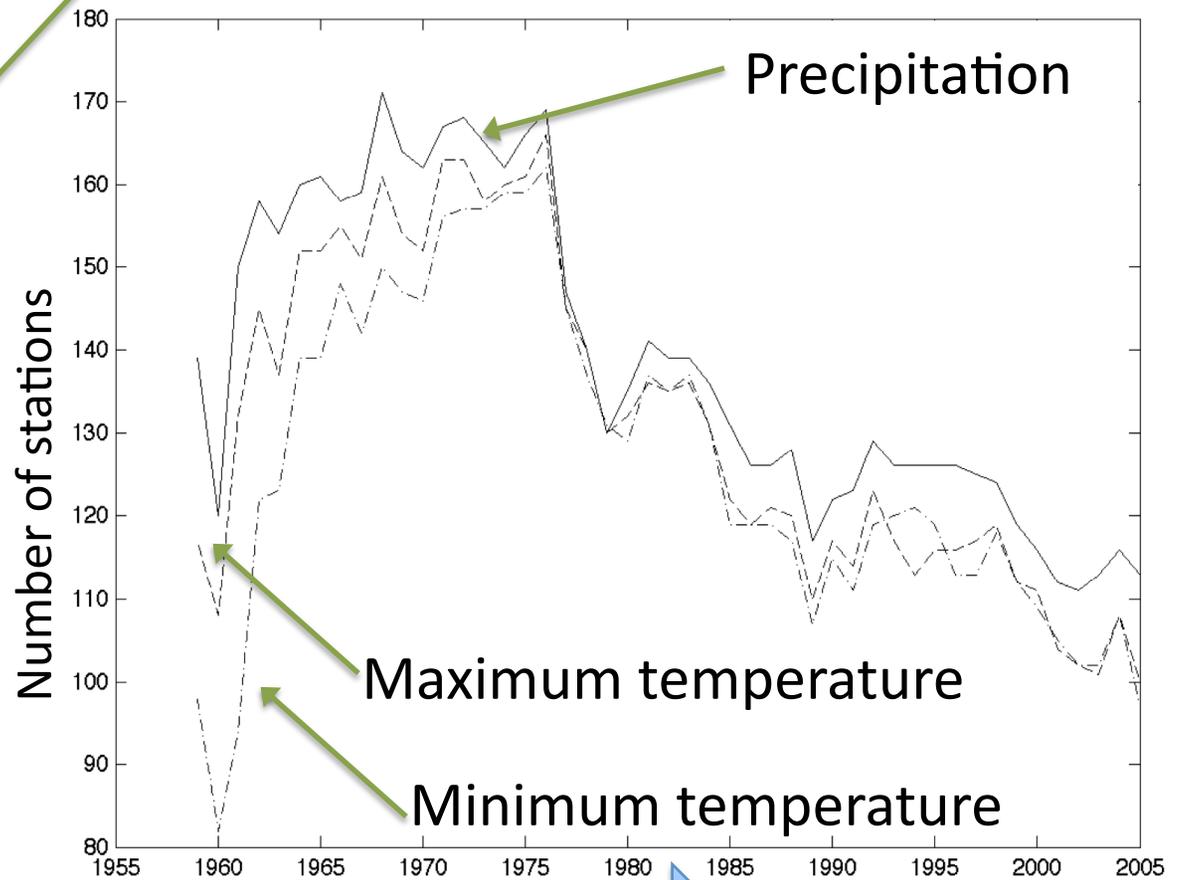
- To develop an error/inhomogeneity detection algorithm for meteorological data.
- To implement the proposed methods in the free open-source (GNU) language R.
- To apply the methods to the CLARIS LPB databases.

The monitoring stations



Dense coverage

Sparse coverage

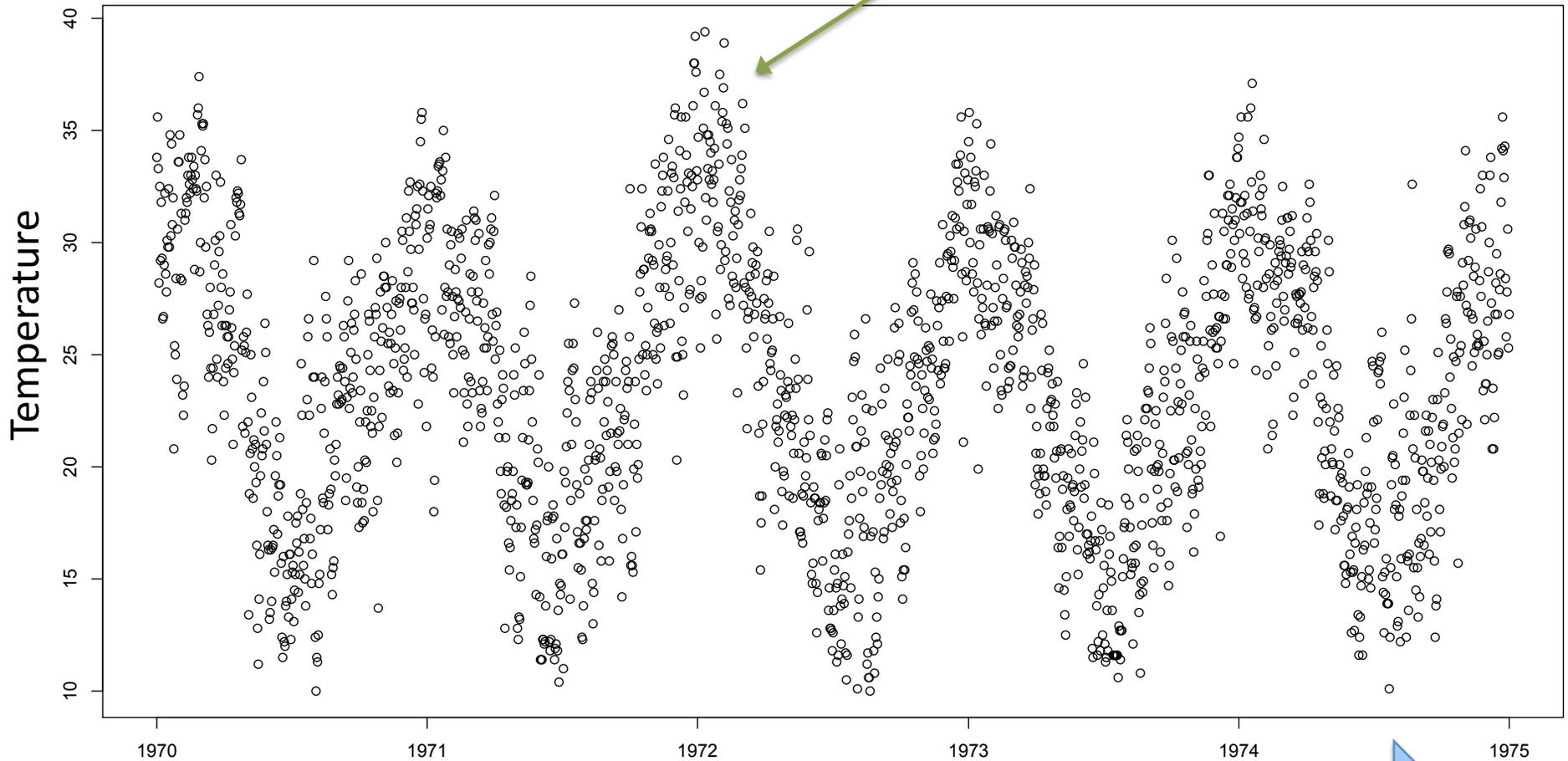


Time

Temperature (maximum) data

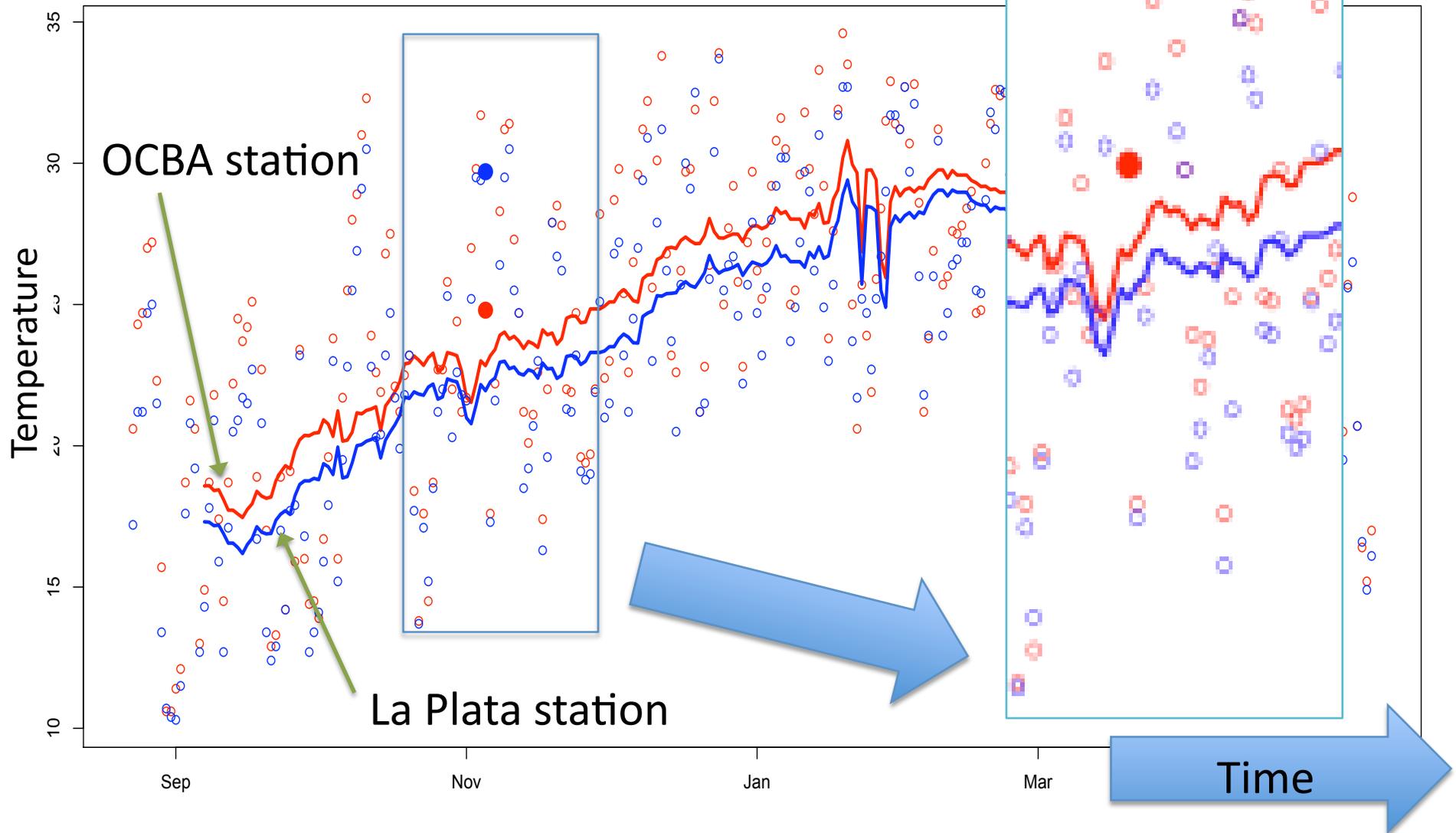
Rosario's maximum temperature series

Warmer summer



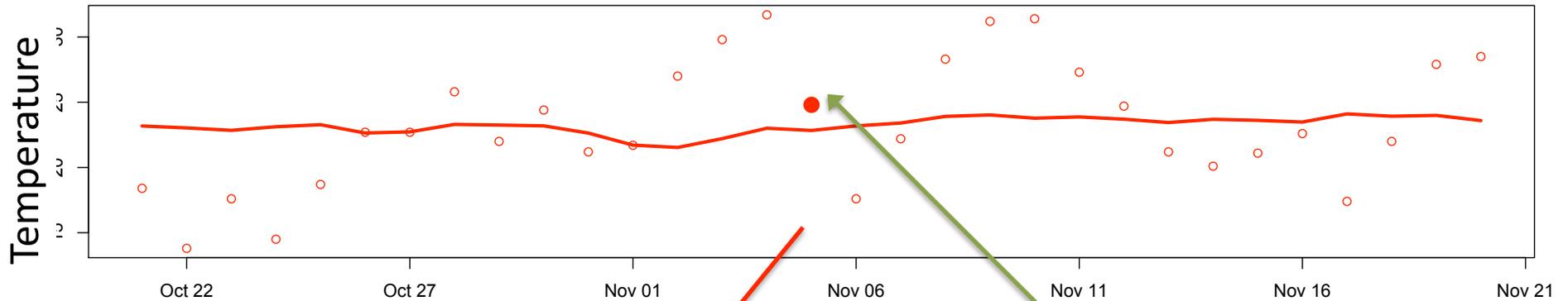
Error detection problem

Evolution of temperatures



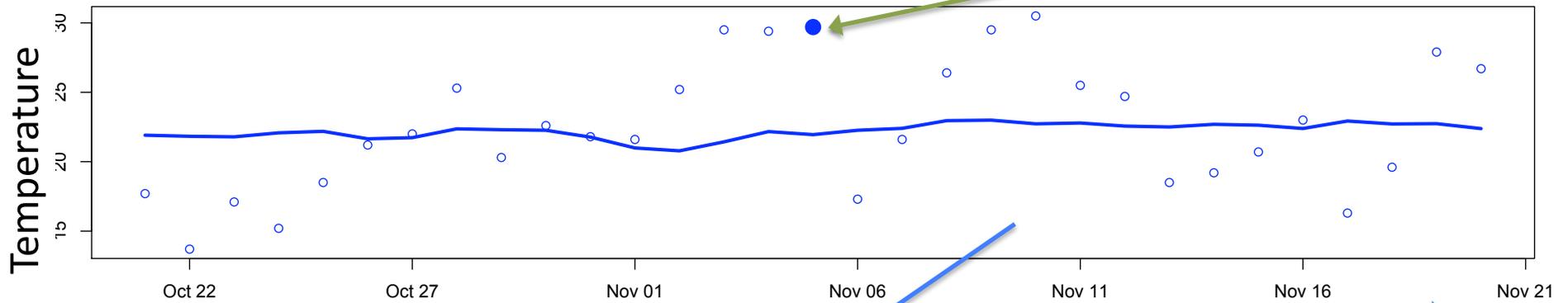
The temperature series

Temperatures within the window



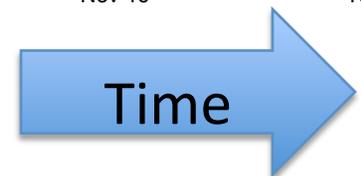
O. C. Buenos Aires station

Suspect values



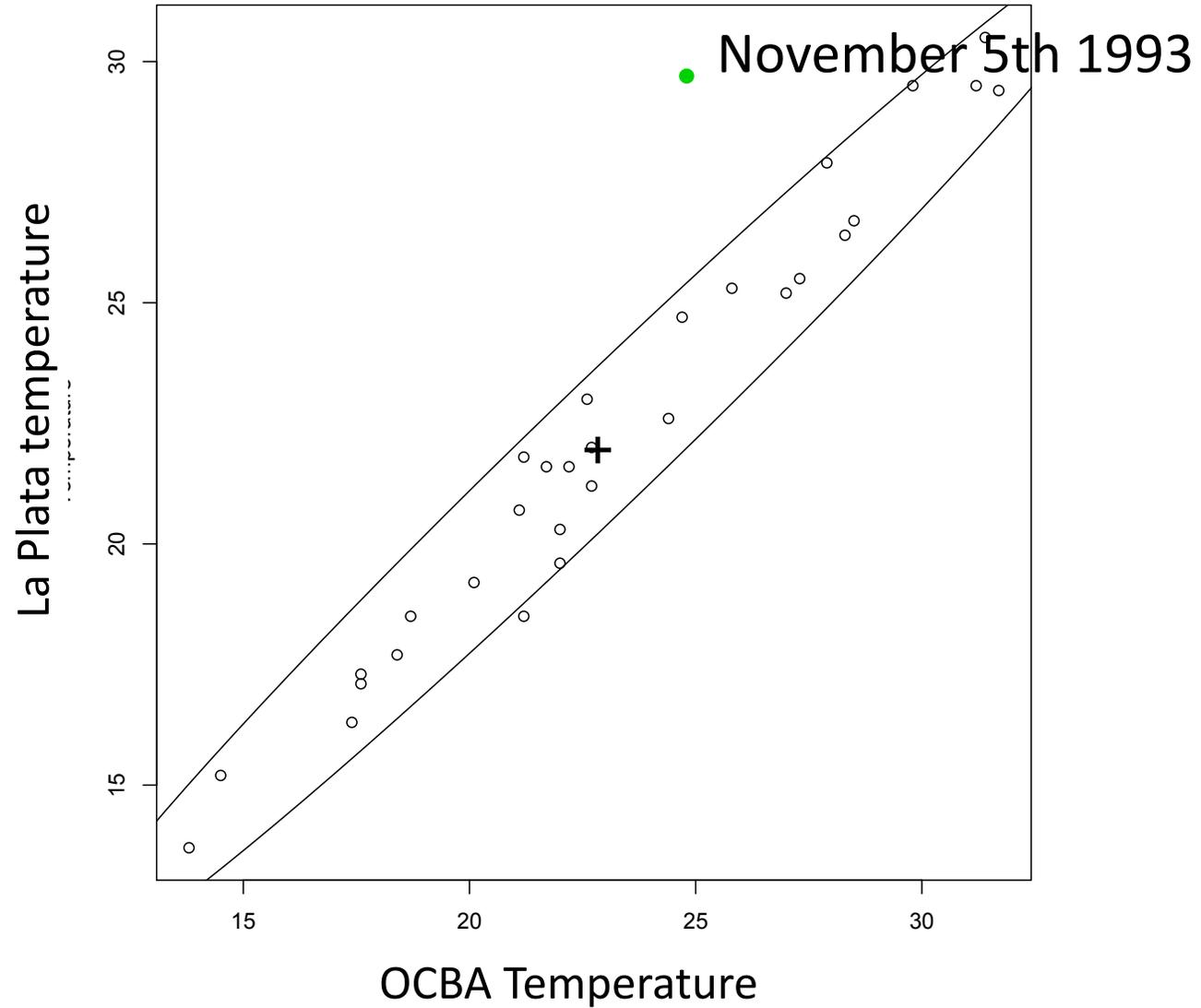
La Plata station

Time

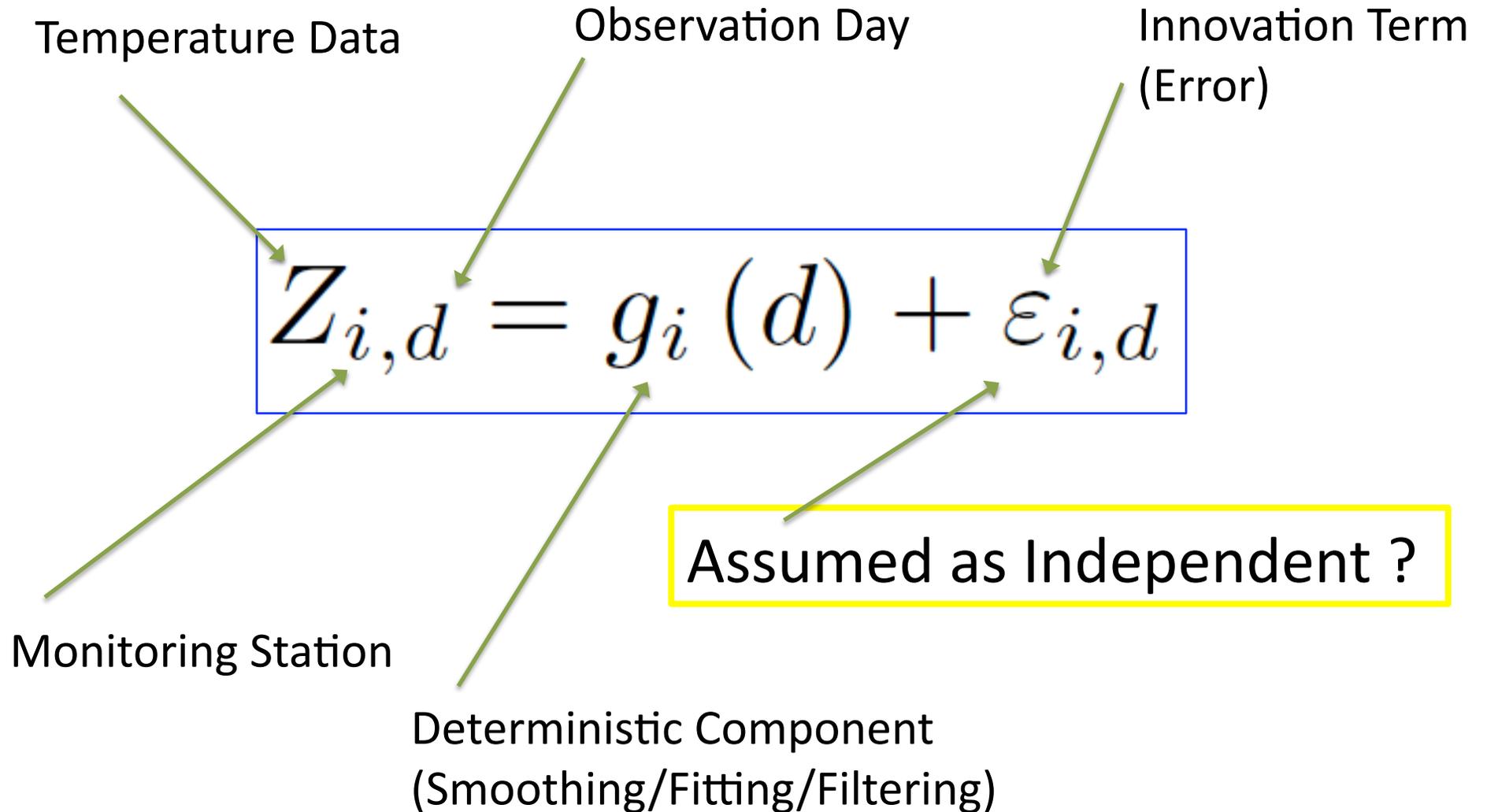


Joint temperature values

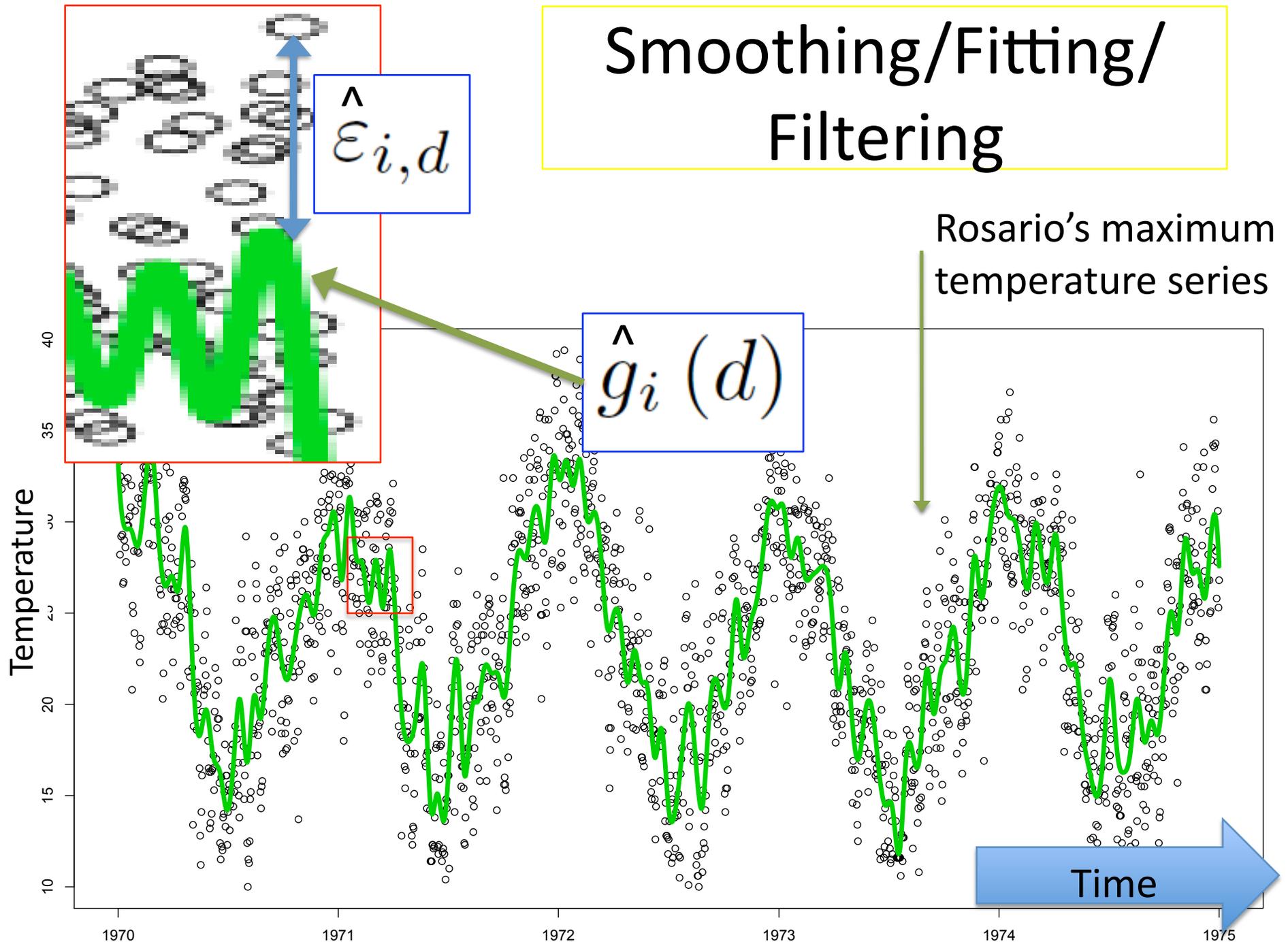
Joint temperatures within the window



The setup



Smoothing/Fitting/ Filtering



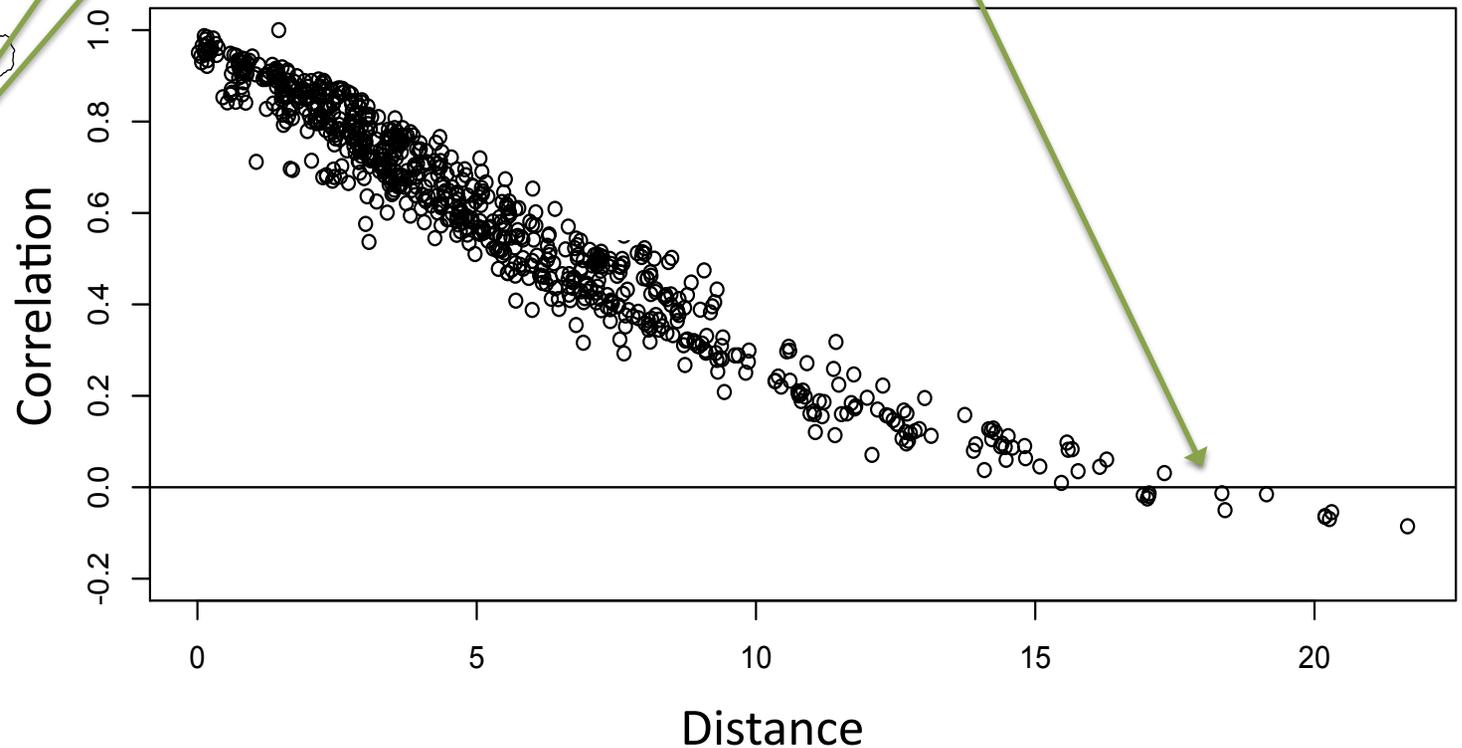
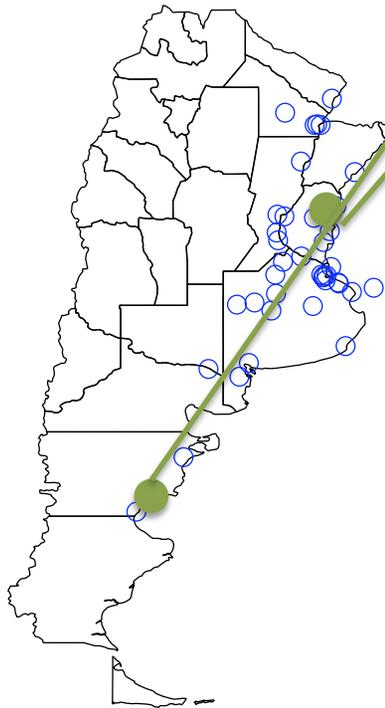
Smoothing/Fitting/Filtering Choices

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

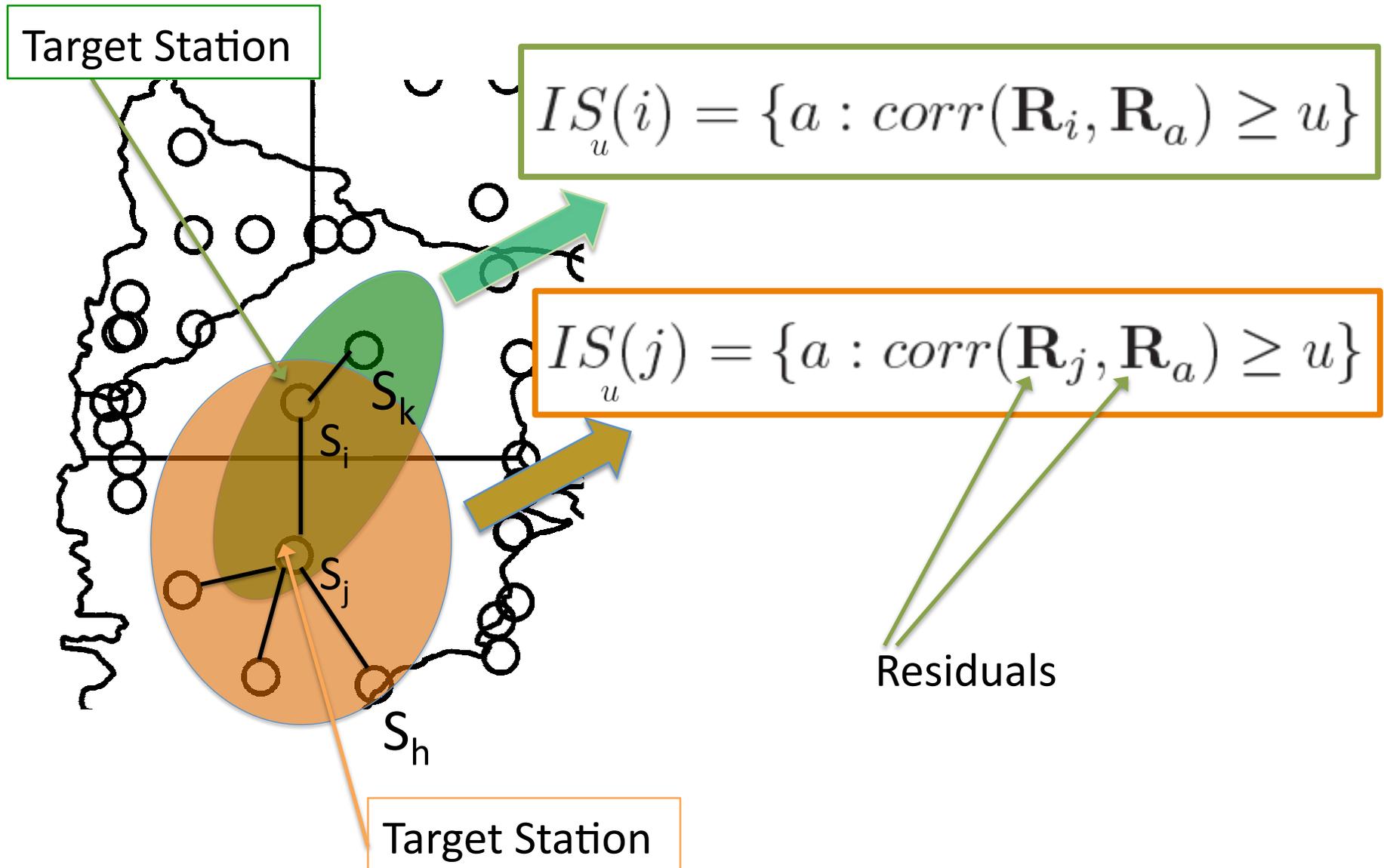
Historical Correlation Matrix of Residuals

		Comodoro	Concordia	
Comodoro		1	0.01	
Concordia		0.01	1	

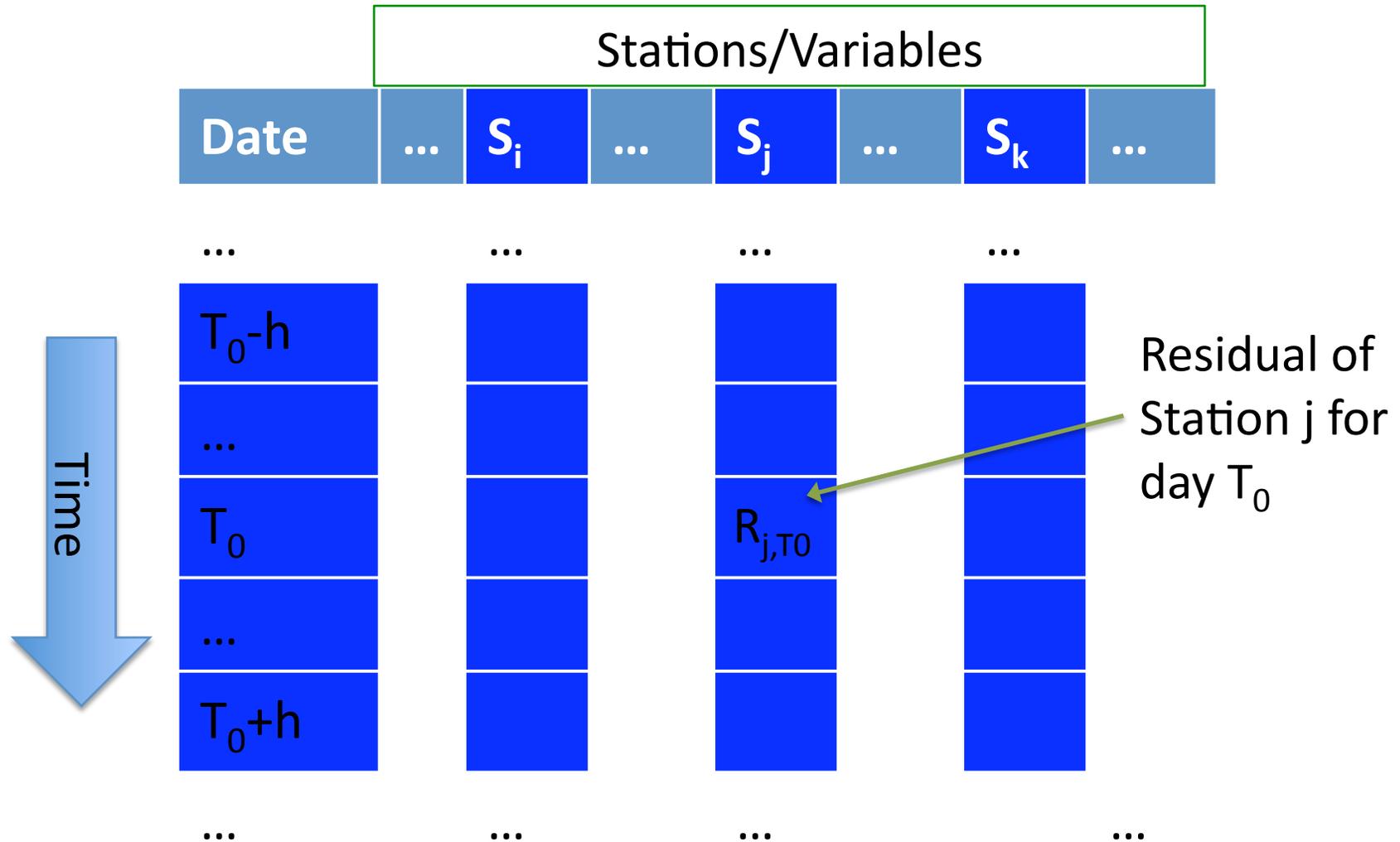
bandwidth = 365



Influence Set of a Target Station



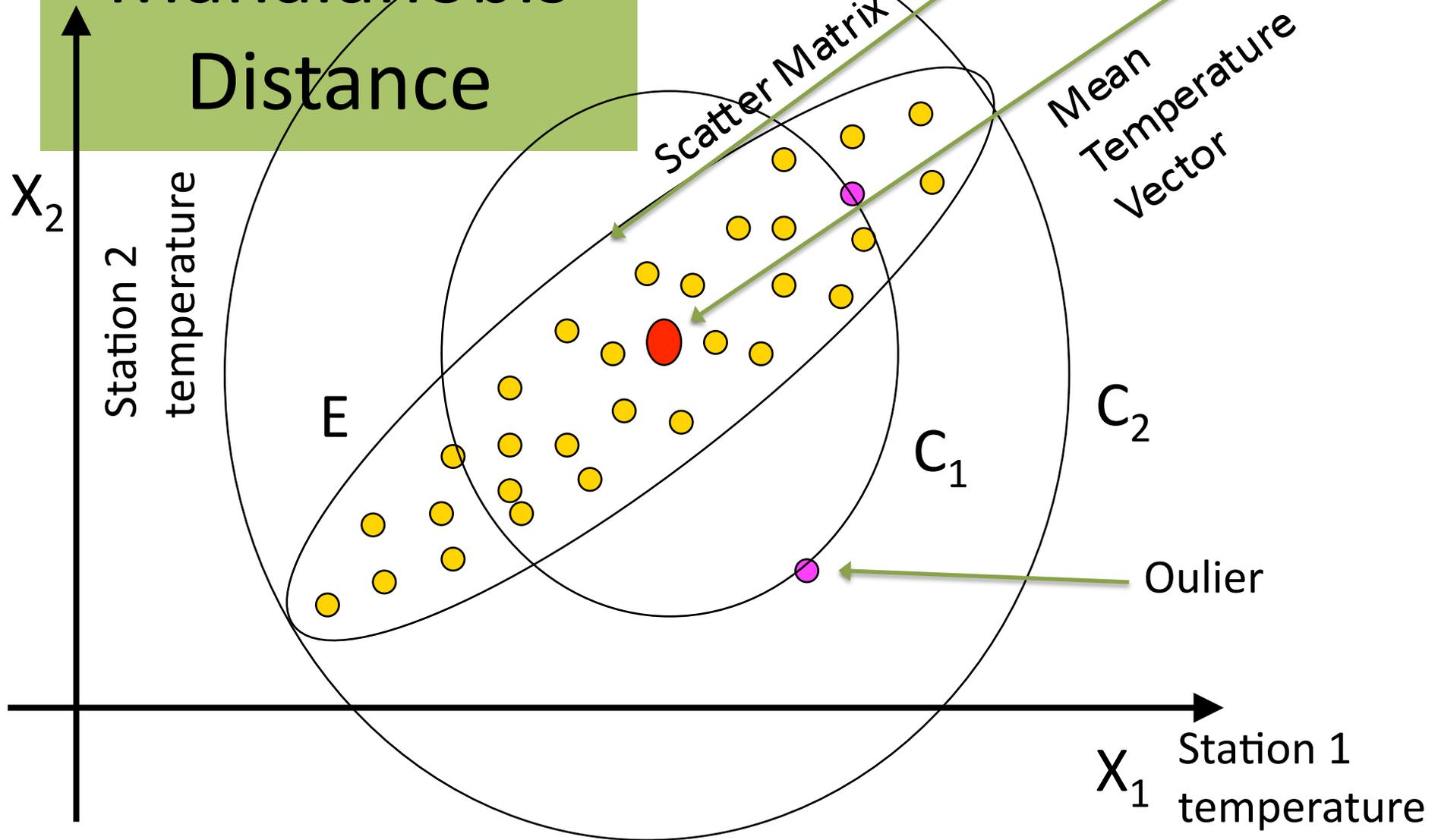
So, for each Influence Set we have:



What is an Outlying Multivariate
Observation?

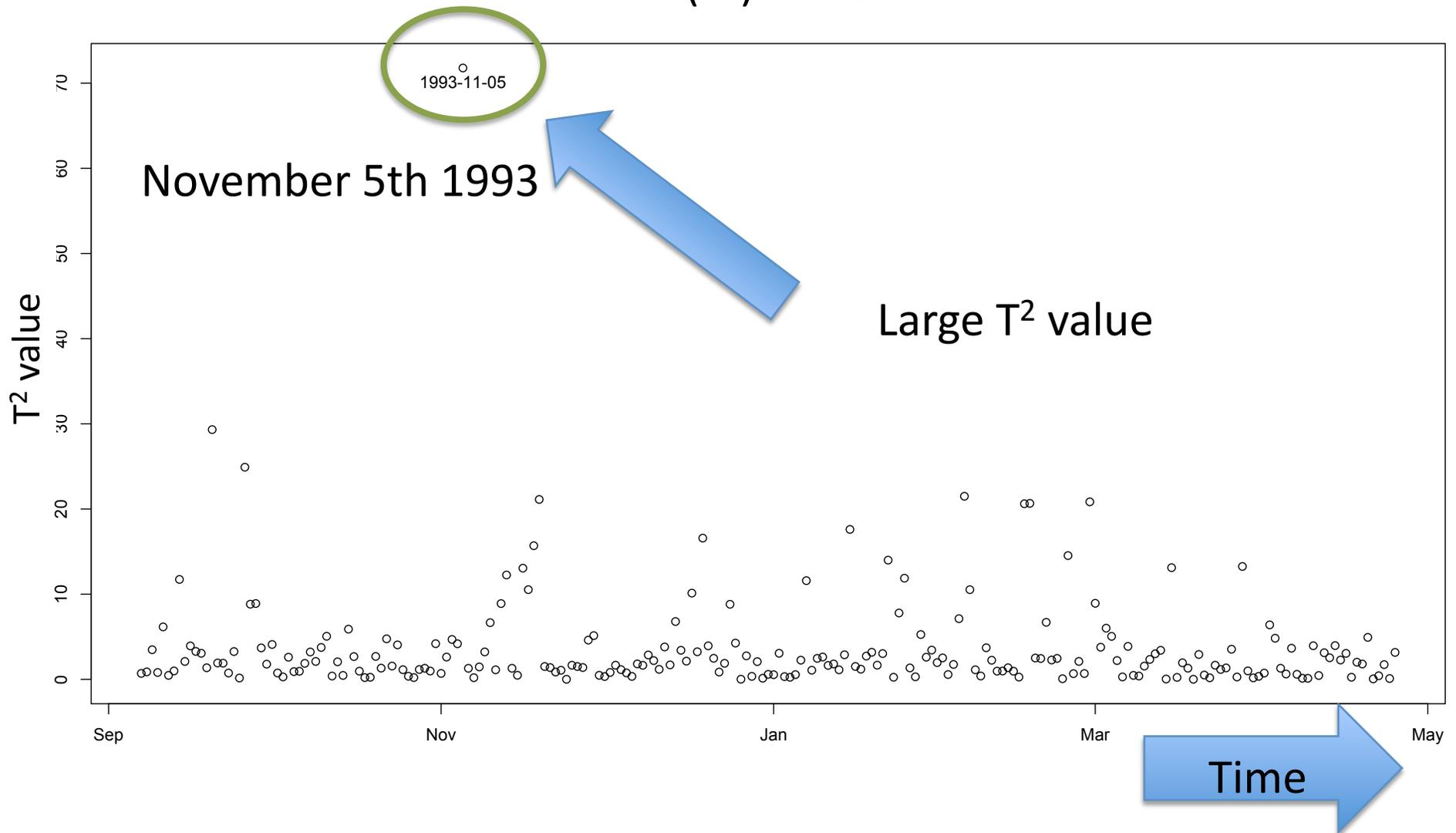
T² Statistic / Mahalanobis Distance

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

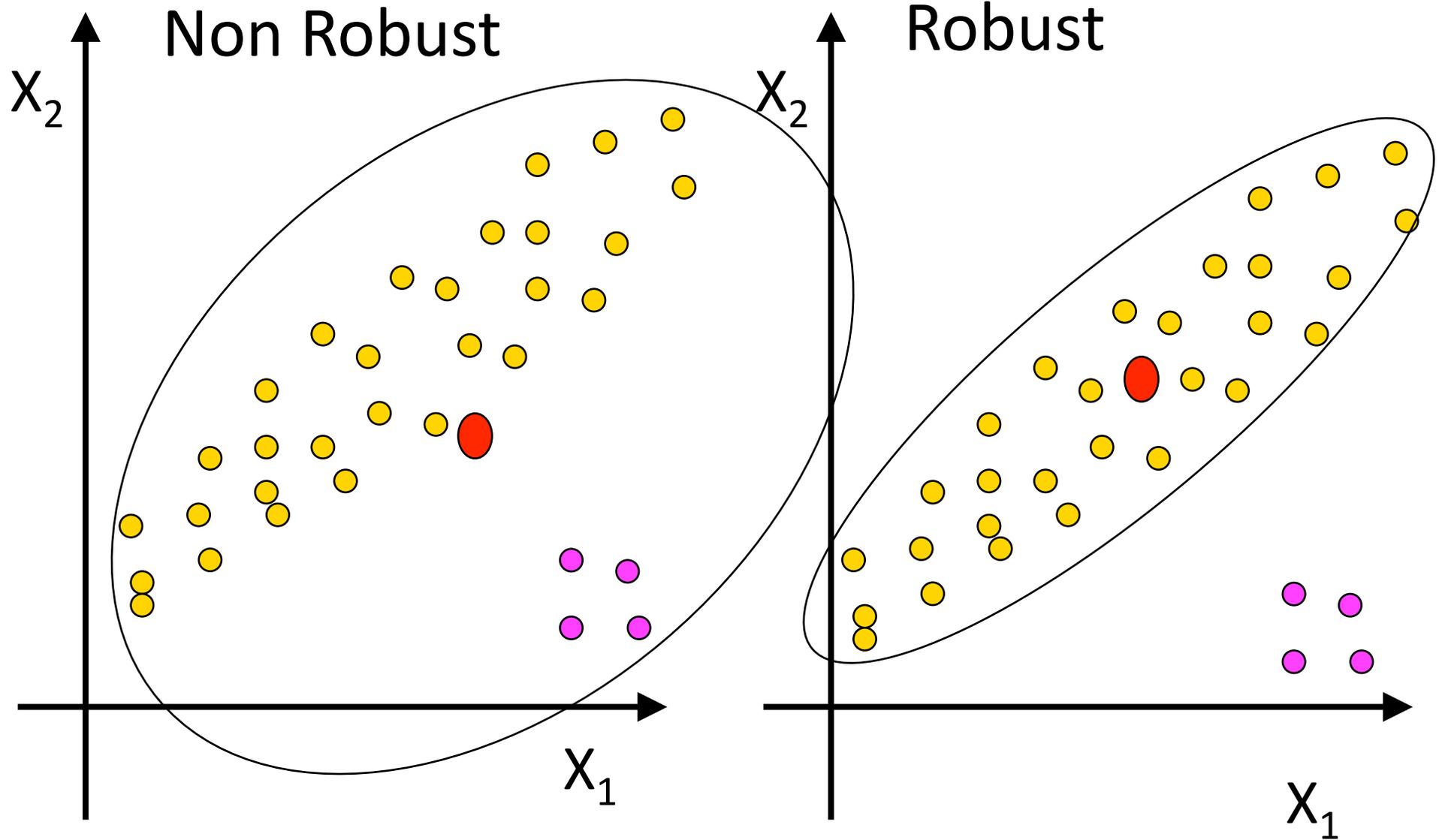


Measuring Outlyingness : La Plata and Buenos Aires

Score (T^2) values



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.

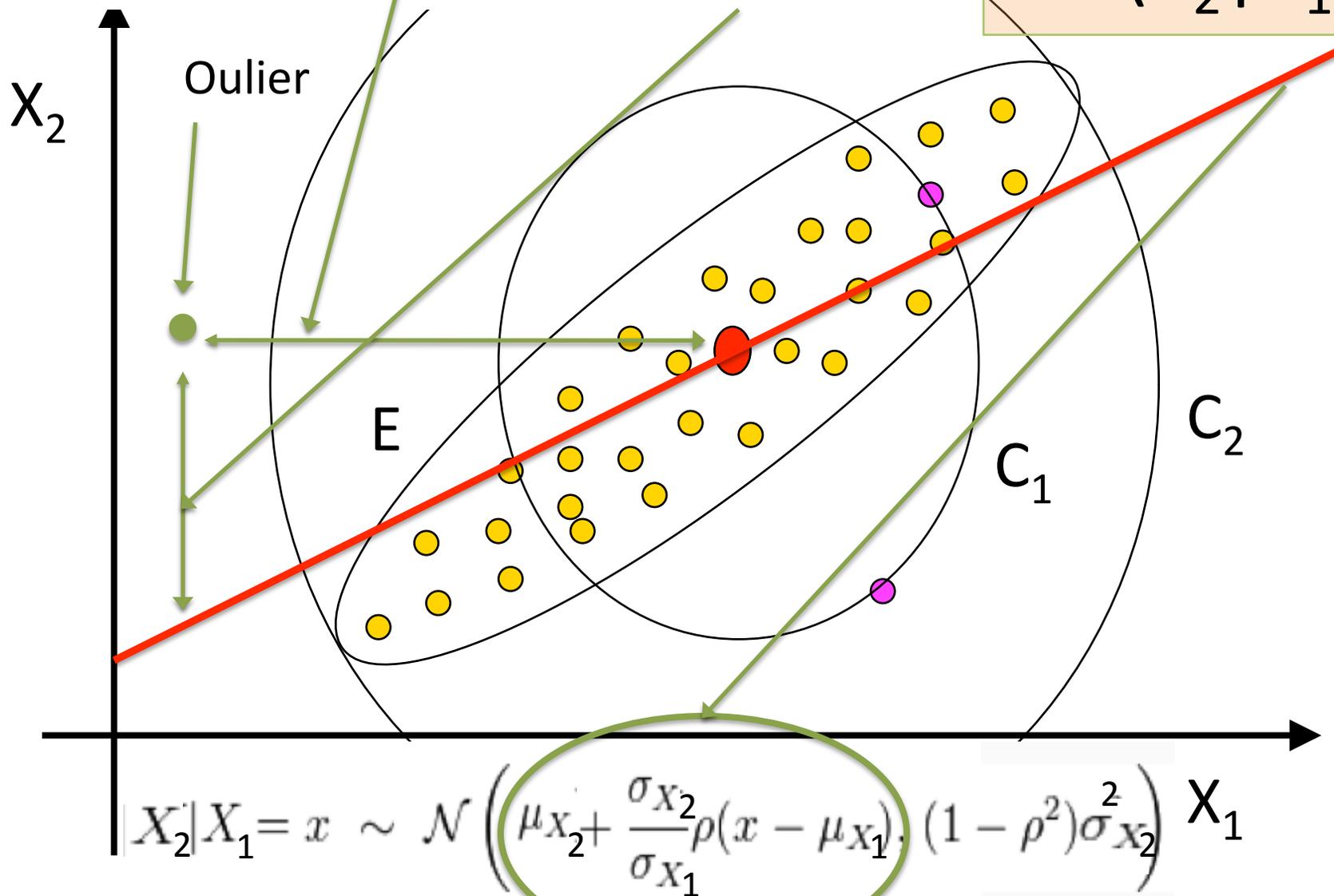
- If the span is too long then the “local (time) effect” could be lost.

- If the span is too short then the method yields unstable estimations ($\# \text{ obs} \gg \# \text{ var}$).

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

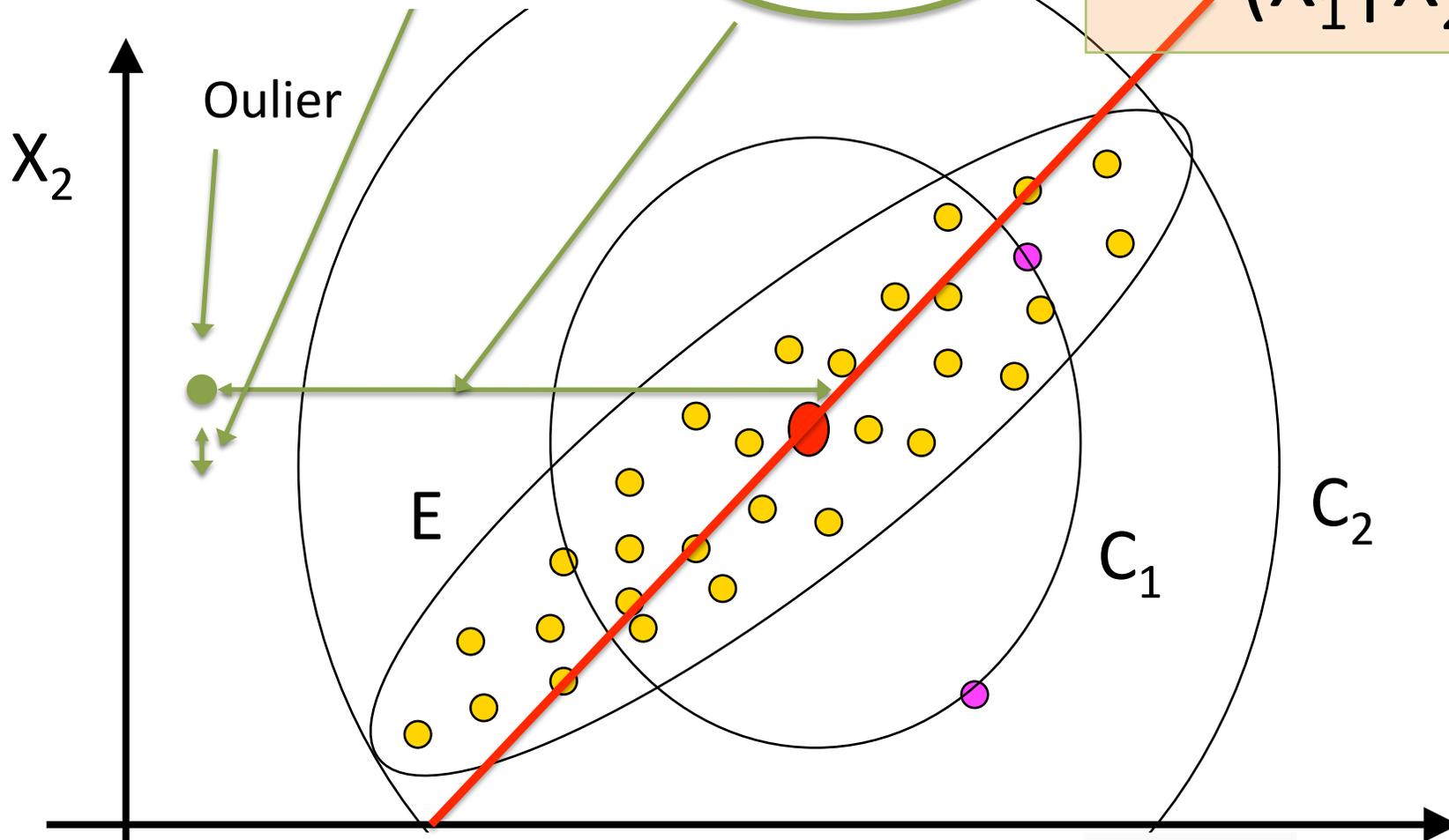
$$T^2 = \frac{(x_1 - \bar{x}_1)^2}{s_1^2} + \frac{(x_2 - \bar{x}_{2.1})^2}{s_{2.1}^2}$$

Blame it to X_2
($X_2 | X_1$)



$$T^2 = \frac{(x_2 - \bar{x}_2)^2}{s_2^2} + \frac{(x_1 - \bar{x}_{1.2})^2}{s_{1.2}^2}$$

Blame it to X_1
 $(X_1 | X_2)$



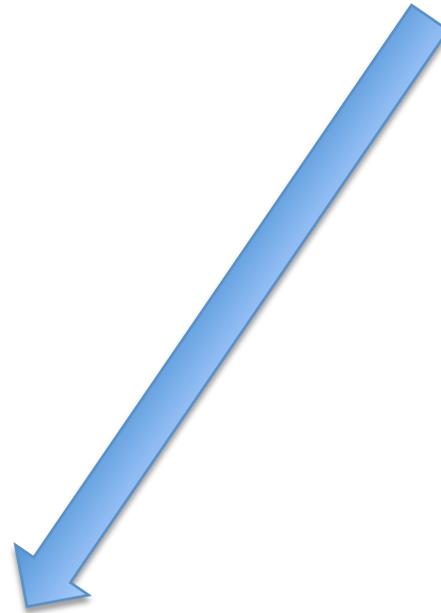
$$X_1 | X_2 = x \sim \mathcal{N} \left(\mu_{X_1} + \frac{\sigma_{X_1}}{\sigma_{X_2}} \rho (x - \mu_{X_2}), (1 - \rho^2) \sigma_{X_1}^2 \right)$$

MYT Decomposition

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



$$T_1^2 = \frac{(x_1 - \bar{x}_1)^2}{s_1^2}$$



Suspect station

$$T_{p/1,2,3,4,\dots,p-1}^2 = \frac{(x_p - \bar{x}_{p/1,2,3,4,\dots,p-1})^2}{s_{p/1,2,3,4,\dots,p-1}^2}$$

The Distance Contribution

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

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

We define the contribution of variable p to the distance as

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

because all terms are positive, the contribution satisfies

$$0 \leq C_p \leq 1$$

Choosing p decompositions (between $p!$ distinct decompositions)

$$T^2 = T_2^2 + T_{3/2}^2 + T_{4/2,3}^2 + T_{5/2,3,4}^2 + \dots + T_{1/2,3,4,\dots,p}^2$$

$$T^2 = T_1^2 + T_{3/1}^2 + T_{4/1,3}^2 + T_{5/1,3,4}^2 + \dots + T_{2/1,3,4,\dots,p}^2$$

$$T^2 = T_1^2 + T_{2/1}^2 + T_{4/1,2}^2 + T_{5/1,2,4}^2 + \dots + T_{3/1,2,4,\dots,p}^2$$

⋮

⋮

⋮

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

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

Error Detection

- If in a certain day the following conditions happen

– T^2 is large

– $C_p = \frac{T_{p/1,2,3,4,\dots,p-1}^2}{T^2}$ is close to 1

then there is an error in station p ,
because all other stations agree

T_{j/A_j}^2 is small

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

Robust Location-Scatter Estimators

- Minimum Covariance Determinant (MCD).
- Minimum Volume Ellipsoid (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 $\mathbf{X}=\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$, 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 \mathbf{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

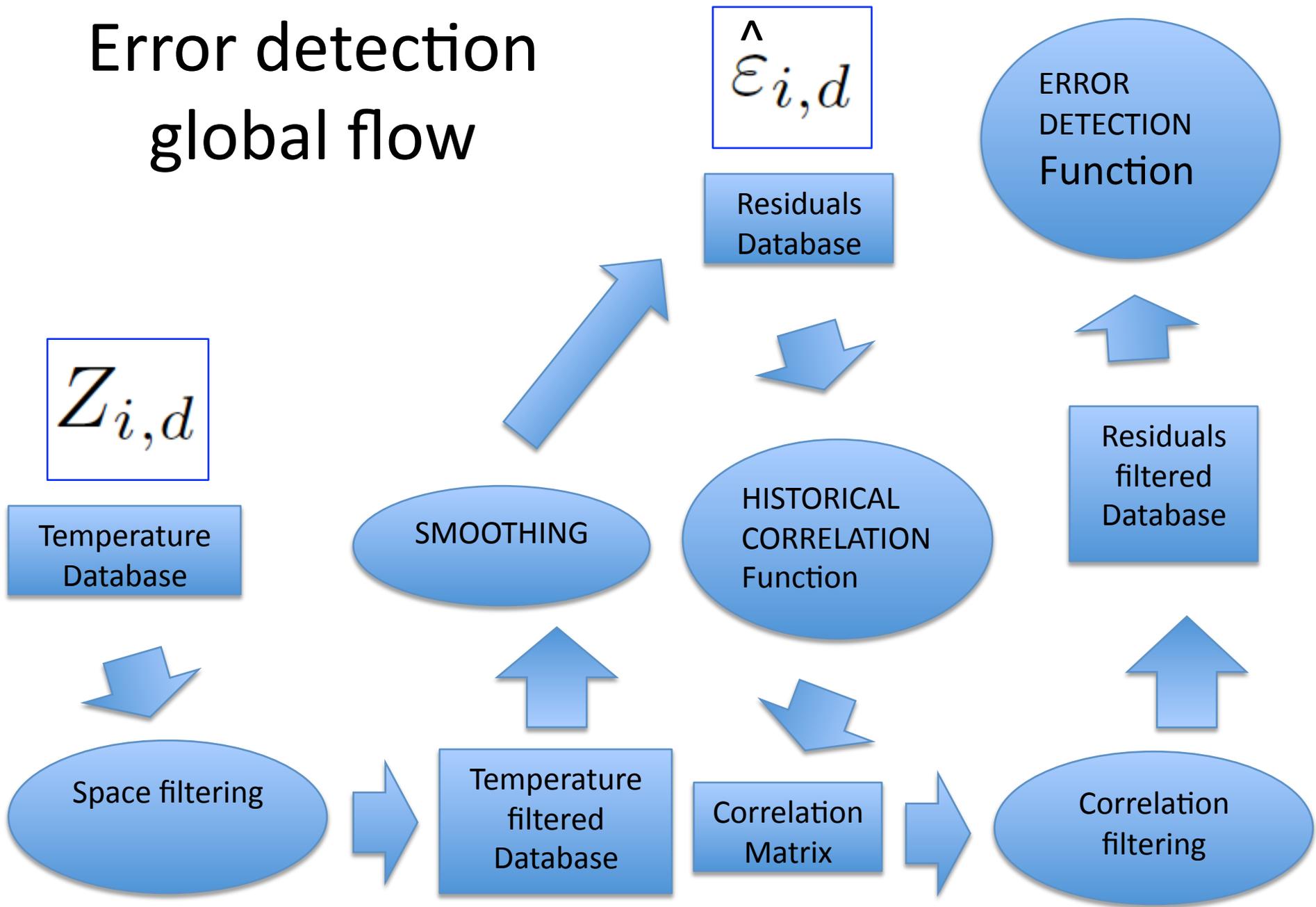
- This estimator relies on the fact that the correlation between monitoring stations do not change over time.

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

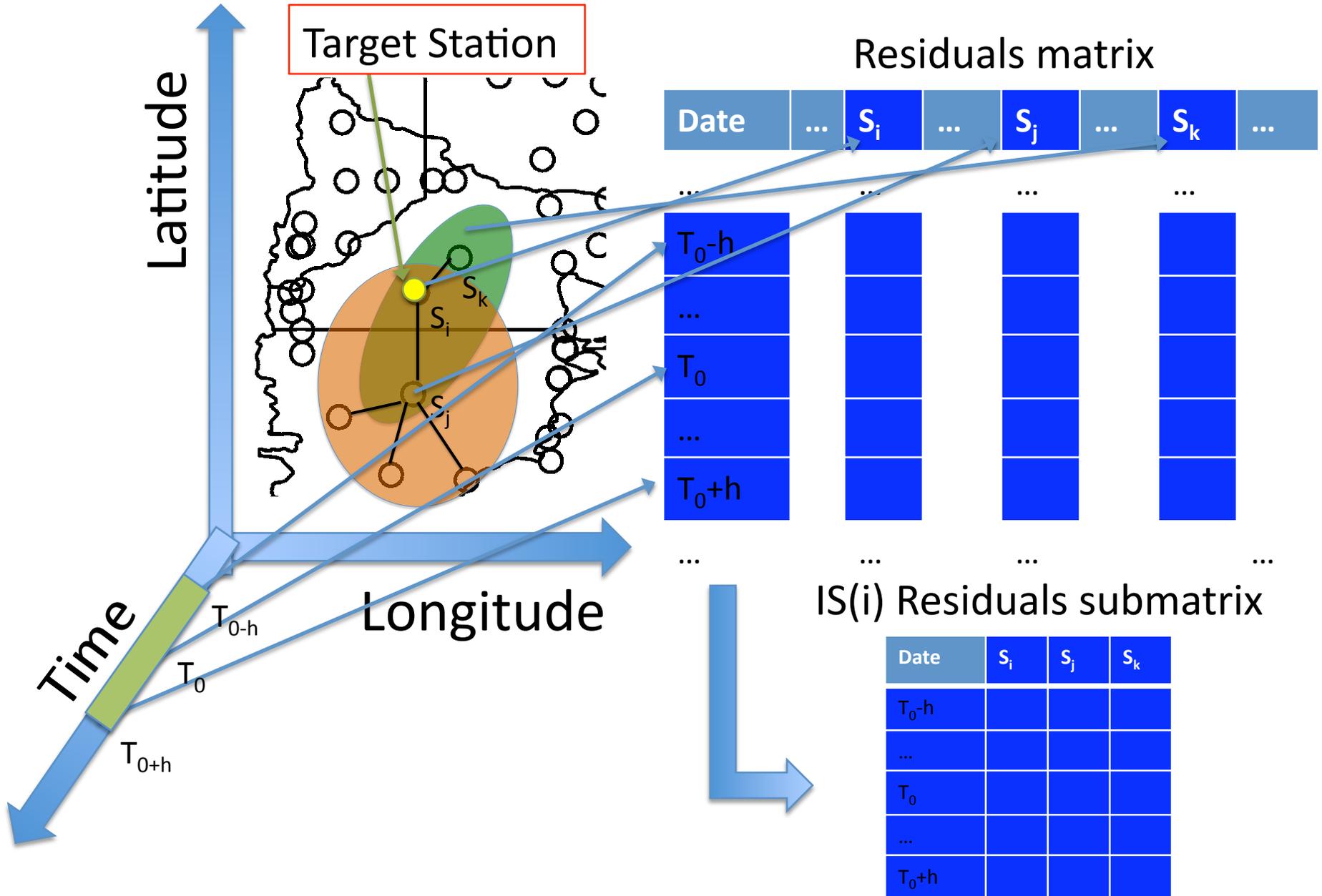
- Thus, to estimate the covariance, only robust estimations of univariate variances are needed.

The Method at Work

Error detection global flow



Error Detection method

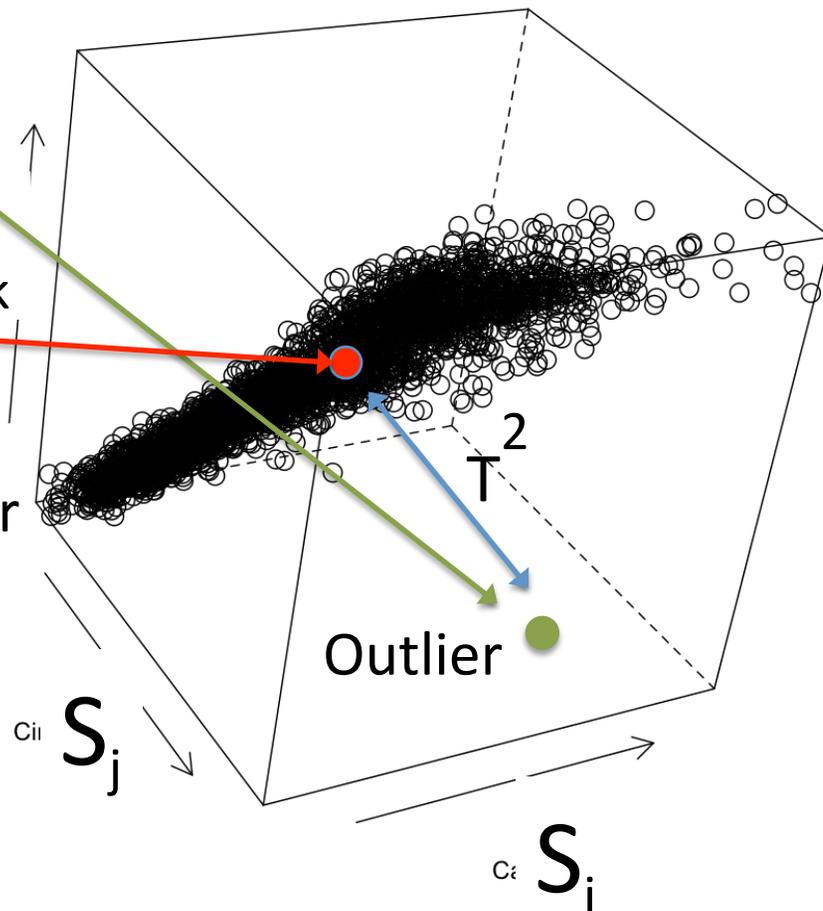


Error Detection method

IS(i) Residuals submatrix

Date	S_i	S_j	S_k
T_{0-h}			
...			
T_0			
...			
T_{0+h}			
Mean			

Residuals scatter plot of stations I, j and k

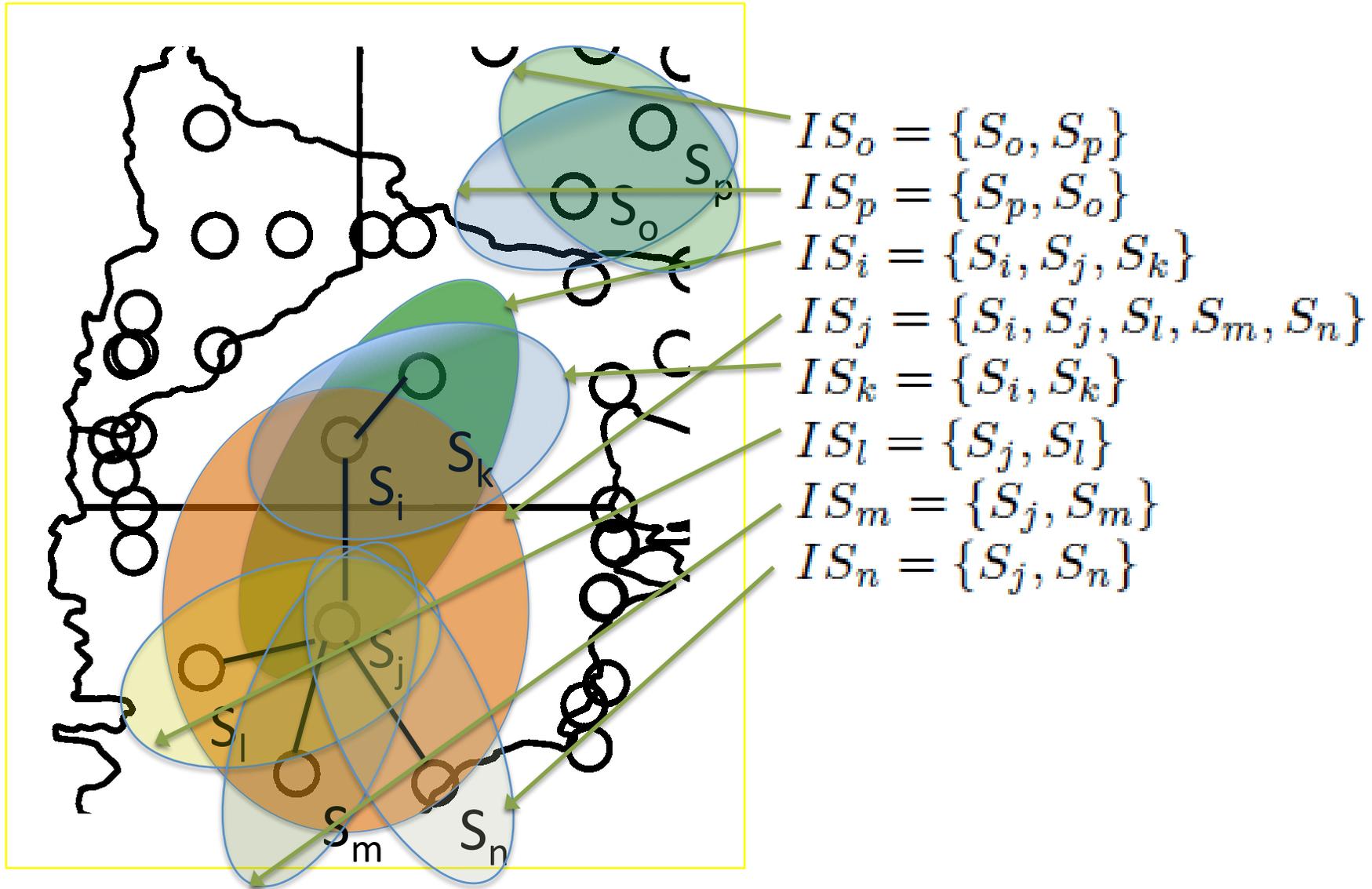


IS(i) Distance Contribution Vector

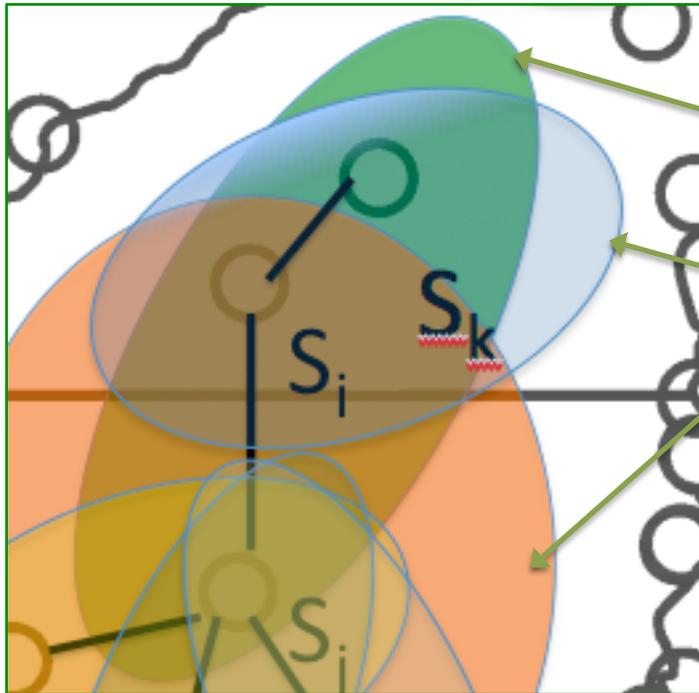
Date	S_i	S_j	S_k
T_0	$T^2_{i.jk}$	$T^2_{j.ik}$	$T^2_{k.ij}$

How to Summarize the
Outlyingness Evidence from
Different Influence Sets ?

A Toy Example (8 stations)



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

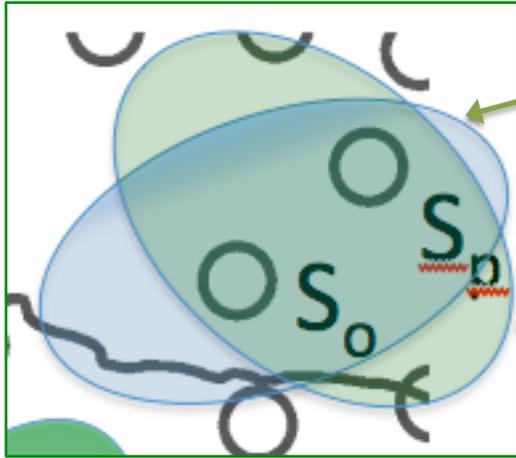
Three different distances are relevant here

$$T_{i/j,k}^2$$

$$T_{i/j,l,m,n}^2$$

$$T_{i/k}^2$$

An Error Occurs in S_o



Both Influence Sets are identical

$$IS_o = \{S_o, S_p\}$$

$$IS_p = \{S_p, S_o\}$$

$$IS_i = \{S_i, S_j, S_k\}$$

$$IS_j = \{S_i, S_j, S_l, S_m, S_n\}$$

$$IS_k = \{S_i, S_k\}$$

$$IS_l = \{S_j, S_l\}$$

$$IS_m = \{S_j, S_m\}$$

$$IS_n = \{S_j, S_n\}$$

Two identical distances are relevant here

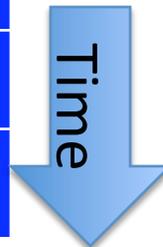
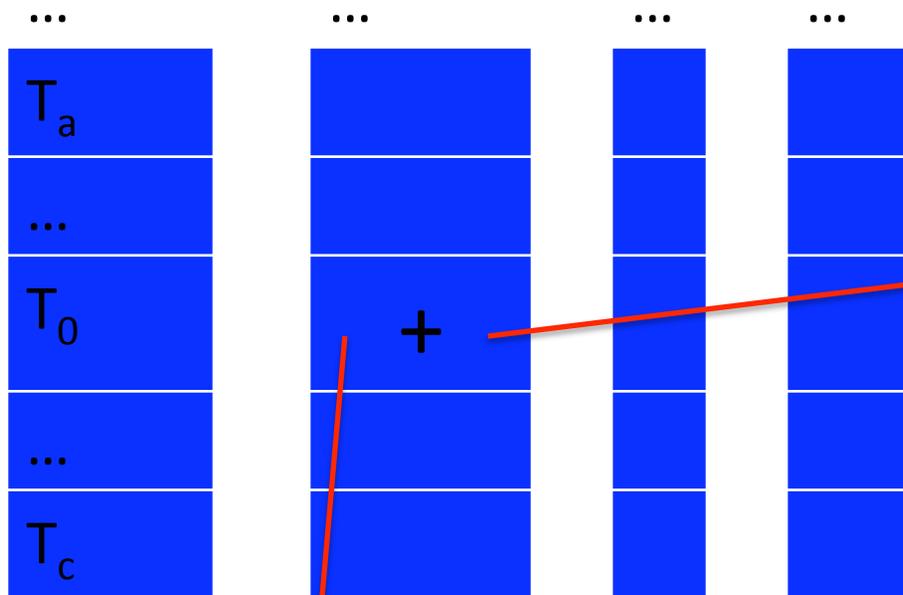
$$T_{o/p}^2$$
$$T_{o/p}^2$$

Distance Contribution Matrix

Date	...	S_i	...	S_j	...	S_k	..
------	-----	-------	-----	-------	-----	-------	----

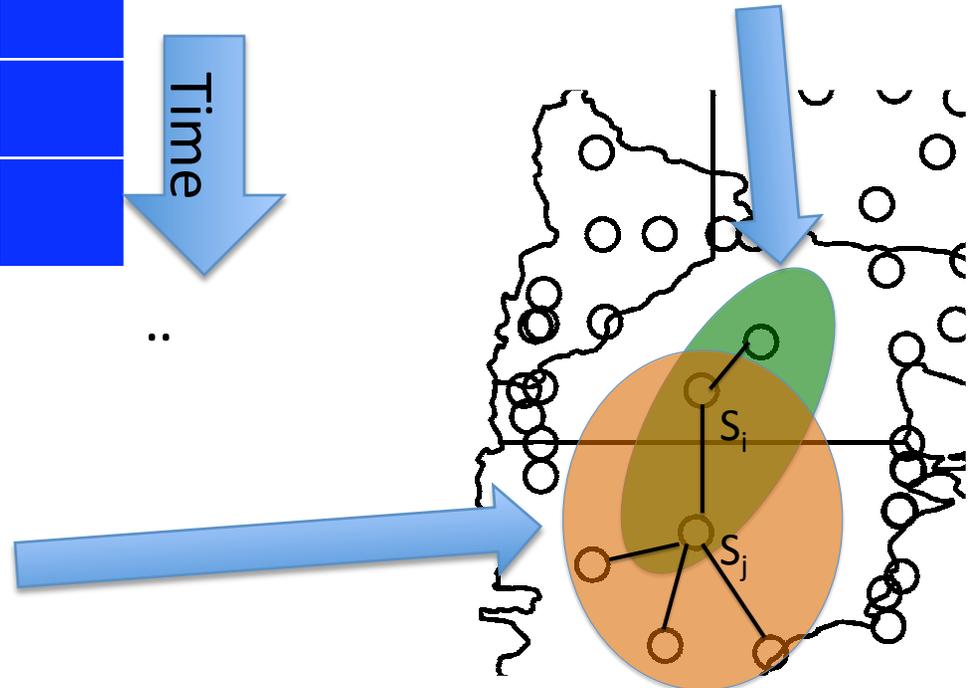
IS(i) Distance Contribution Vector

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



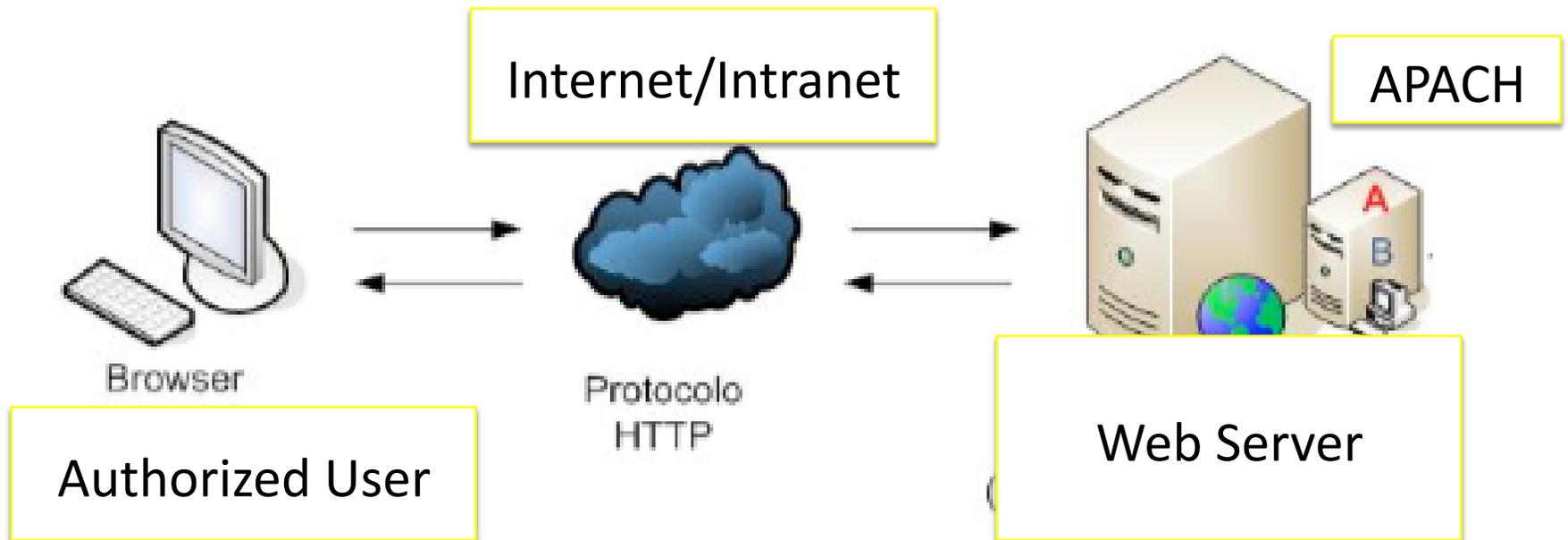
IS(j) Distance Contribution Vector

Date	S_i	S_j	...
T_0	$T^2_{i,j...}$	$T^2_{j,i...}$	$T^2_{...}$



The Algorithms Implementation

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



Uploading a database

The screenshot shows the Apache Daily Climate Data Homogenization interface. At the top left is the Apache logo and the text "Apache Daily Climate Data Homogenization". Below this is a navigation bar with three tabs: "DATABASES", "STATIONS", and "QCP". The "DATABASES" tab is selected, and a yellow arrow points to the "Measurements Databases list" section. This section contains a table with the following structure:

					INPUT	
					Name	Description
					BASEGRAN	60 estaciones en 50 años
					BASEMED	50 estaciones en 30 años

The "BASEGRAN" entry in the table is circled in green, and a red box labeled "Selected database" has an arrow pointing to it.

Executing a Quality Control Process



Selected database



Input data for Error Detection Process

Description: ROSARIO AERO 300 KM 0.8 TMAX

Select the process to execute

Control de consistencia de informacio (DE 1)

Select the Databases to process

	NAME	Description	State	VISIBI
<input checked="" type="checkbox"/>	BASEGRAN	60 estaciones en 50 anios	Completado	Privated
<input type="checkbox"/>	BASEMED	50 estaciones en 30 anios	Completado	Privated
<input type="checkbox"/>	BASERED	BASERED	Completado	Privated

QCP: Parameters determination

Input data for Error Detection Process

Description • ROSARIO AERO 300 KM 0.8 TMAX|

Select the process to execute

Control de consistencia de informacio (DE 1)

Select the Databases to process

	NAME	Description	State	VISIBI
<input checked="" type="checkbox"/>	BASEGRAN	60 estaciones en 50 años	Completado	Privated
<input type="checkbox"/>	BASEMED	50 estaciones en 30 años	Completado	Privated
<input type="checkbox"/>	BASERED	BASERED	Completado	Privated

Begin 1900-01-01
Formato AAAA-MM-DD

End 2010-01-01
Formato AAAA-MM-DD

Measurements > Temperatura
Prec
Rad
Wind

Value Maximum

Doubtful 100

Minimal Correlation 0.8

Target variable

Time span of analysis

Correlation Threshold

Selecting the stations

The image shows a map of Argentina with various stations marked. A yellow circle highlights the Rosario area, which is labeled as the 'Target station'. A red box labeled 'Unselected station' points to a station in the north. A red box labeled 'Reference station' points to a station in the east. A red box labeled 'Target station' points to the Rosario area. A yellow arrow points from the 'Reference station' box to the 'Target station' box. A table on the right lists station details, with the 'AEROPARQUE AERO (AR) - SMN 87582' station checked.

Radio de Seleccion (km)		300	
Estacion	Latitud	Longit	
<input type="checkbox"/>	BARILOCHE AERO (AR) - 87765	-41.1	
<input type="checkbox"/>	COMODORO RIVADAVIA AERO (AR) - SMN 87860	-45.78	
<input type="checkbox"/>	MAQUINCHAO (AR) - SMN 87774	-41.2	
<input type="checkbox"/>	PUERTO DESEADO AERO (AR) - SMN 87896	-47.73	
<input type="checkbox"/>	TRELEW AERO (AR) - SMN 87828	-43.2	
<input checked="" type="checkbox"/>	AEROPARQUE AERO (AR) - SMN 87582	-34.3	
<input type="checkbox"/>	AZUL AERO I (AR) - SMN 87642	-36.4	
<input type="checkbox"/>	BAHIA BLANCA AERO (AR) -	-38.4	

Executing a Quality Control Process

Error Detection Data			
Reference	ROSARIO AERO 300 KM 0.8 TMAX		
User	edgardo		
Station	ROSARIO AERO		
Process	Selected Process		
	Process	NAME	Version
	 Control de consistencia de informacio	DE	1
	Seleccion de mediciones para ejecucion de proceso	SELECCION	1
Database	Selected Databases		
	Database	Description	
	BASEGRAN	60 estaciones en 50 anios	
Station	ROSARIO AERO AEROPARQUE AERO BUENOS AIRES EL PALOMAR AERO EZEIZA AERO GUALEGUAYCHU AERO JUNIN AERO MARCOS JUAREZ AERO		

View button (points to the process icon)

Stations involved (points to the station list)

Mahalanobis Distance (T^2)

List of potential errors



Apache
Daily Climate Data
Homogenization

Suspect date

of stations

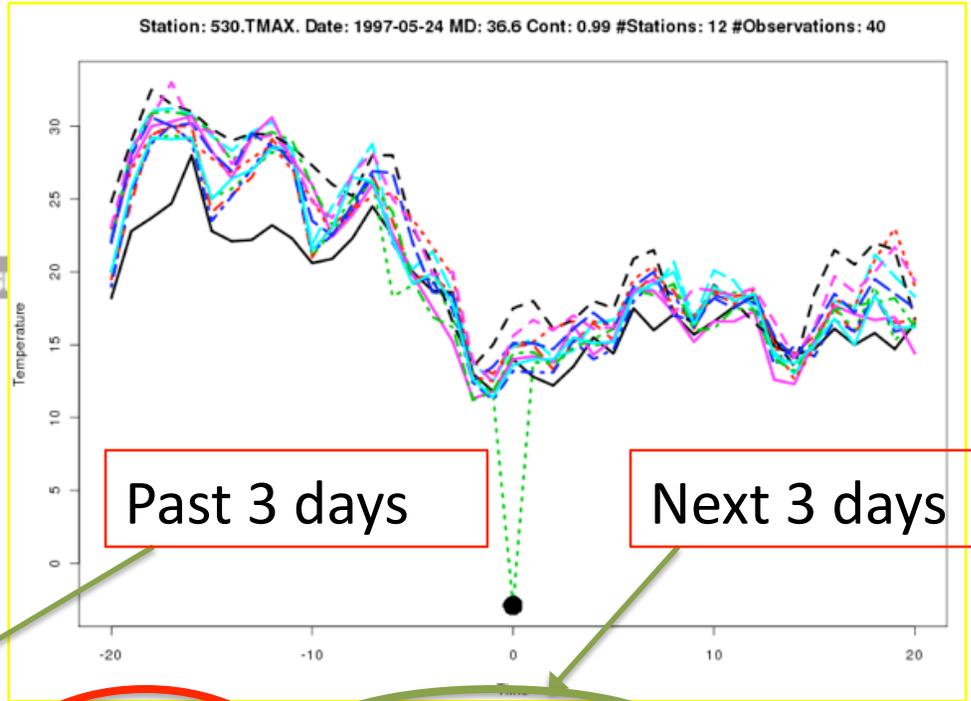
Observed Results Measurements					
Order by		Severity		Max Contribution	
Severity	MD	Max Cont.	Date	Observed	Station
1	44.047	0.991	1955-07-04	41	6
2	36.598	0.992	1997-05-24	40	12
3	30.545	0.943	1974-05-25	41	12
4	30.264	0.921	1956-02-19	29	8
5	27.809	0.938	1953-10-31	35	5
6	26.895	0.926	1984-11-04	41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

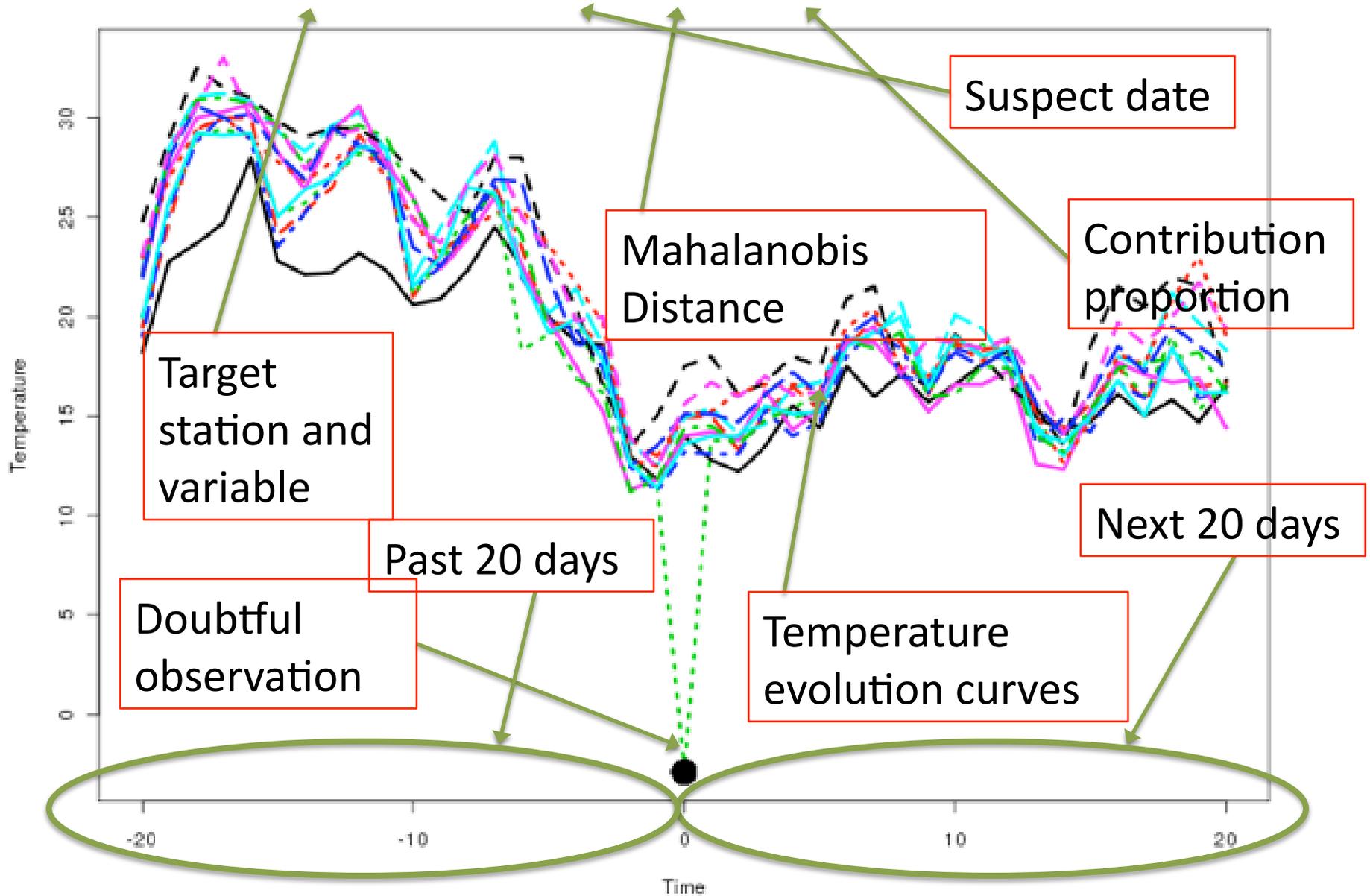


Station	Cont		-3	-2	-1	1997-05-24	+1	+2	+3
EL PALOMAR AERO (-34.36 , -58.36 , 1)	0.992	Min	14.8	4.8	0.8	13.9	3.5	-0.3	3.2
		Max	16.800	12.900	11.600	-2.900	13.800	13.700	16.100
SAN MIGUEL (-34.33 , -58.44 , 1)	0.183	Min	12	6	1.5	0.4	3.6	2.8	6.6
		Max	17.800	12.600	11.400	13.600	14.000	14.000	15.500
EZEIZA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

Tmin and Tmax mixed up

The Graphical Output in detail

Station: 530.TMAX. Date: 1997-05-24 MD: 36.6 Cont: 0.99 #Stations: 12 #Observations: 40

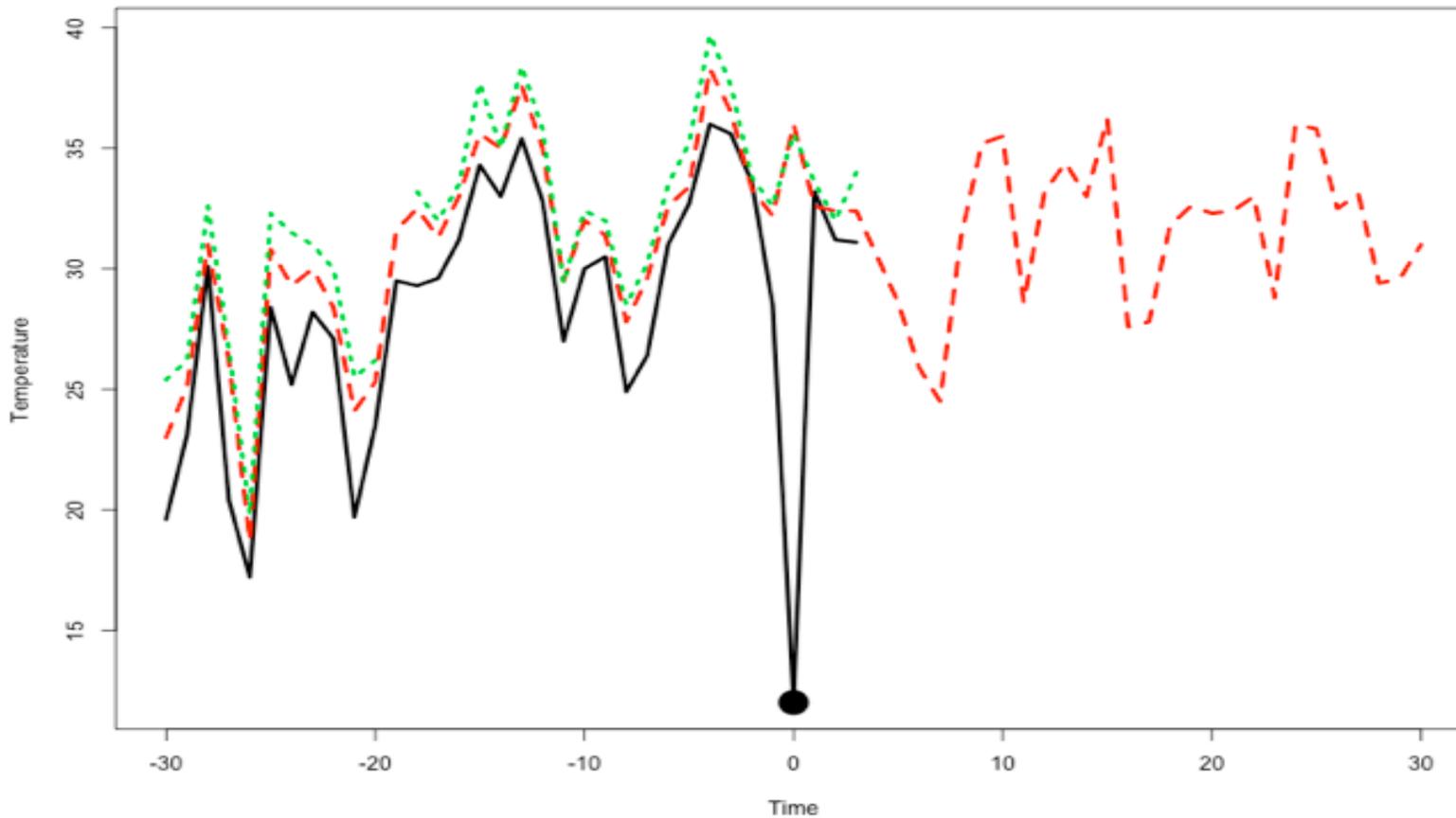


Some Examples

Univariate Outlier (3 Stations)

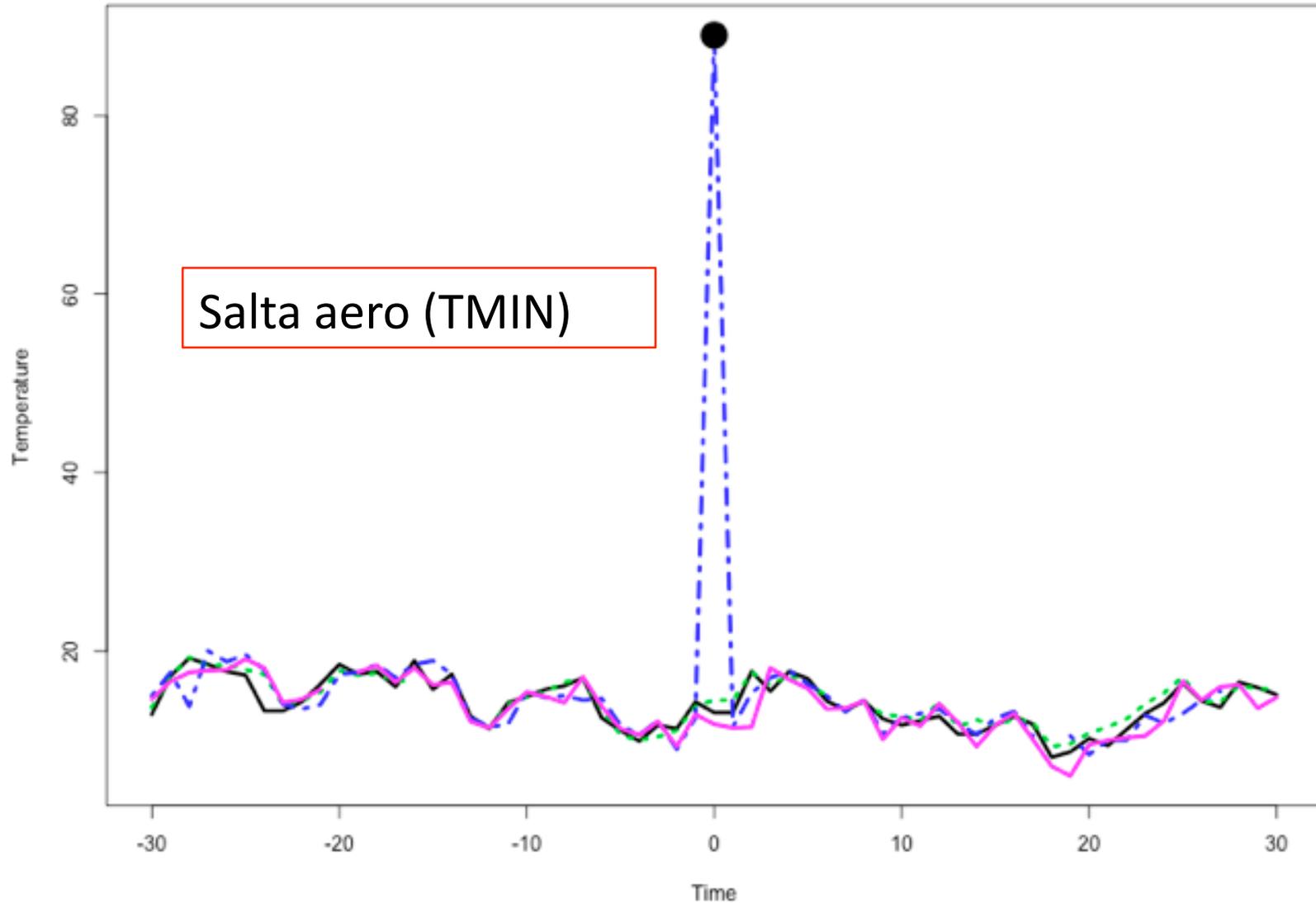


Station: 87711.TMAX. Date: 2001-12-28 Dist: 38.54 Contr: 0.97 #Stations: 3 #Observations: 33

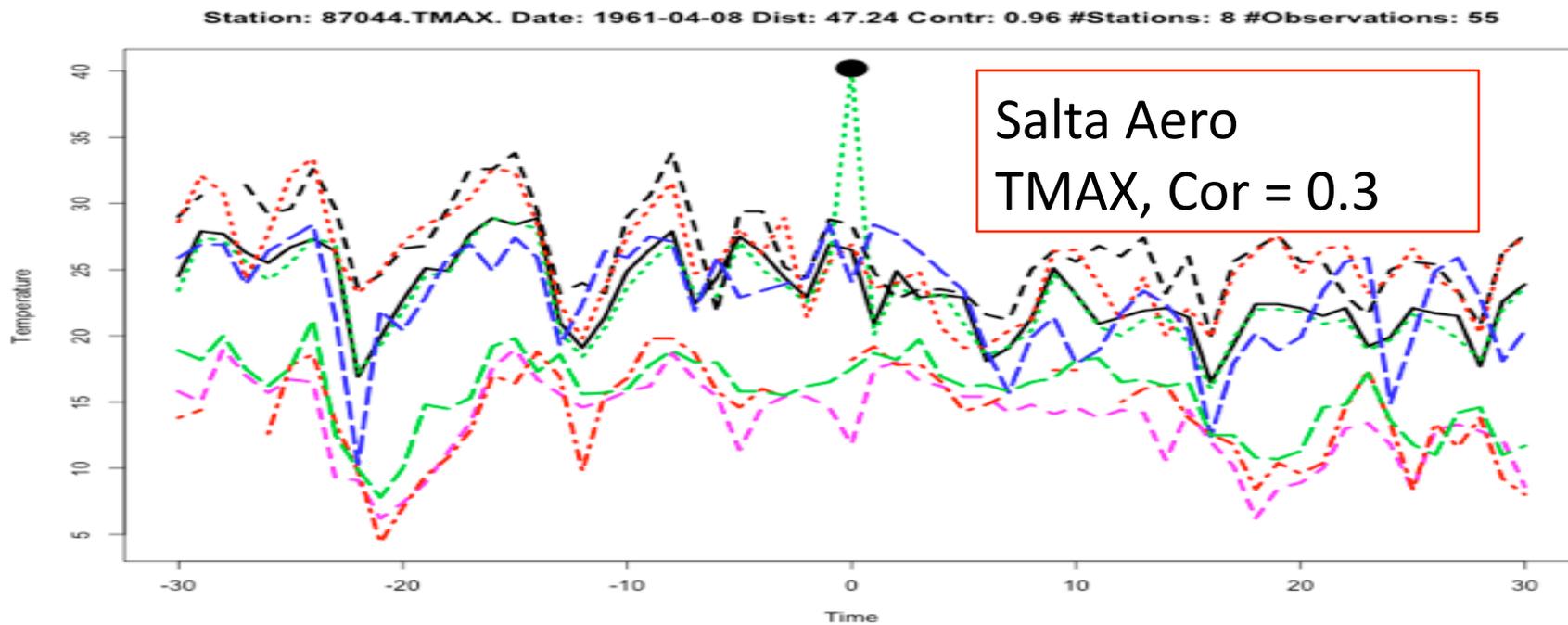
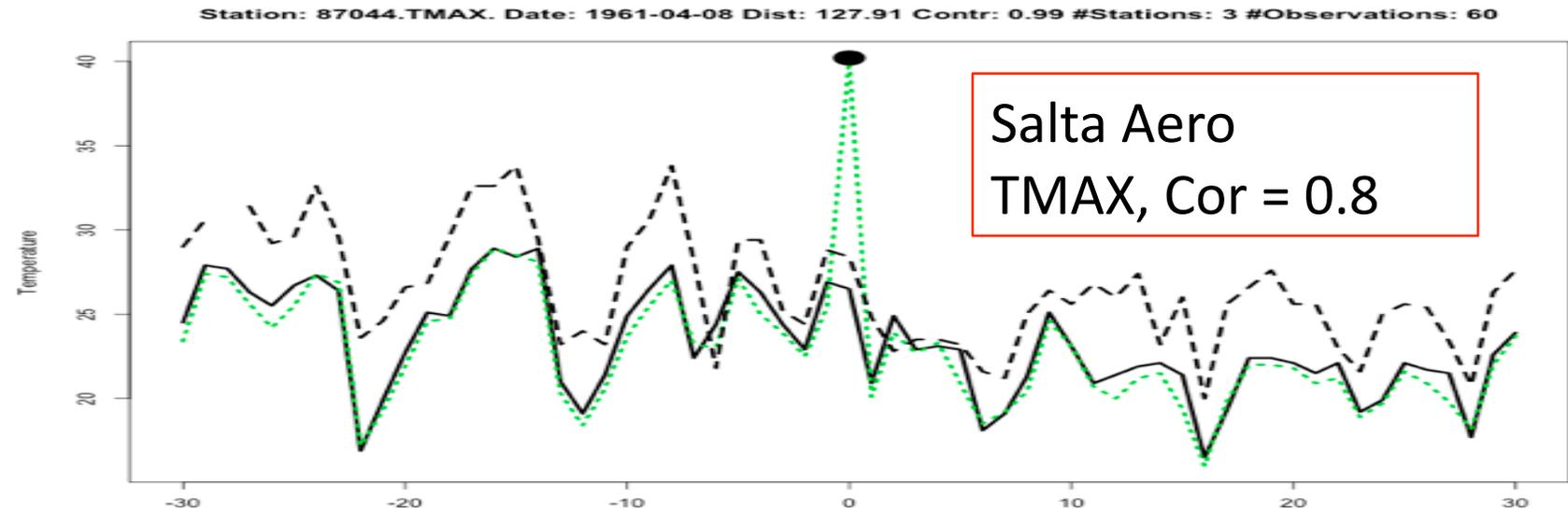


Huge Univariate Outlier

Station: 87045.TMIN. Date: 1963-04-01 Dist: 3105.03 Contr: 1 #Stations: 4 #Observations: 54

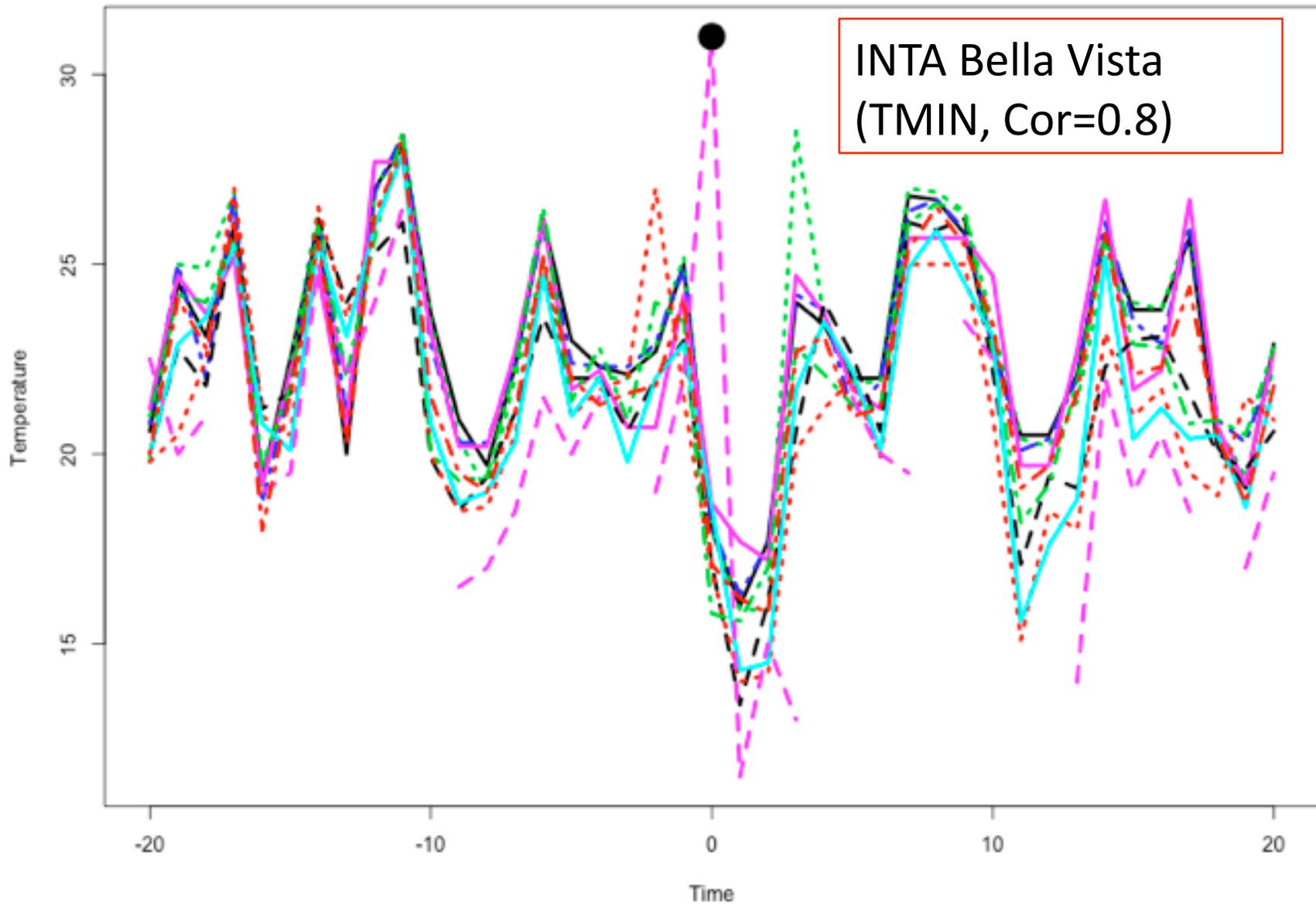


Changing the Correlation



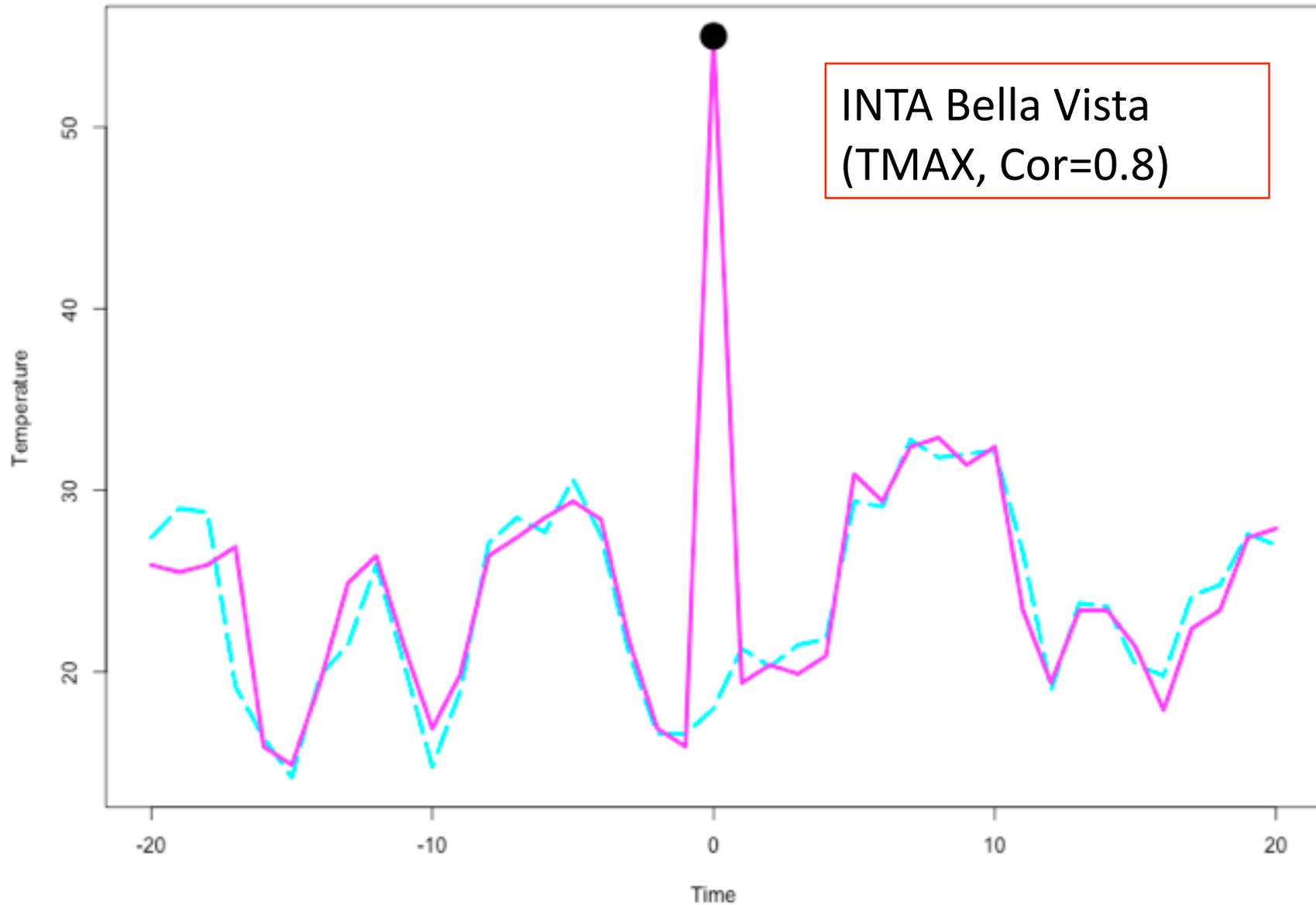
Potential Delay in Temperature Change

Station: 87286.TMIN. Date: 1973-01-24 Dist: 210.17 Contr: 0.98 #Stations: 10 #Observations: 33



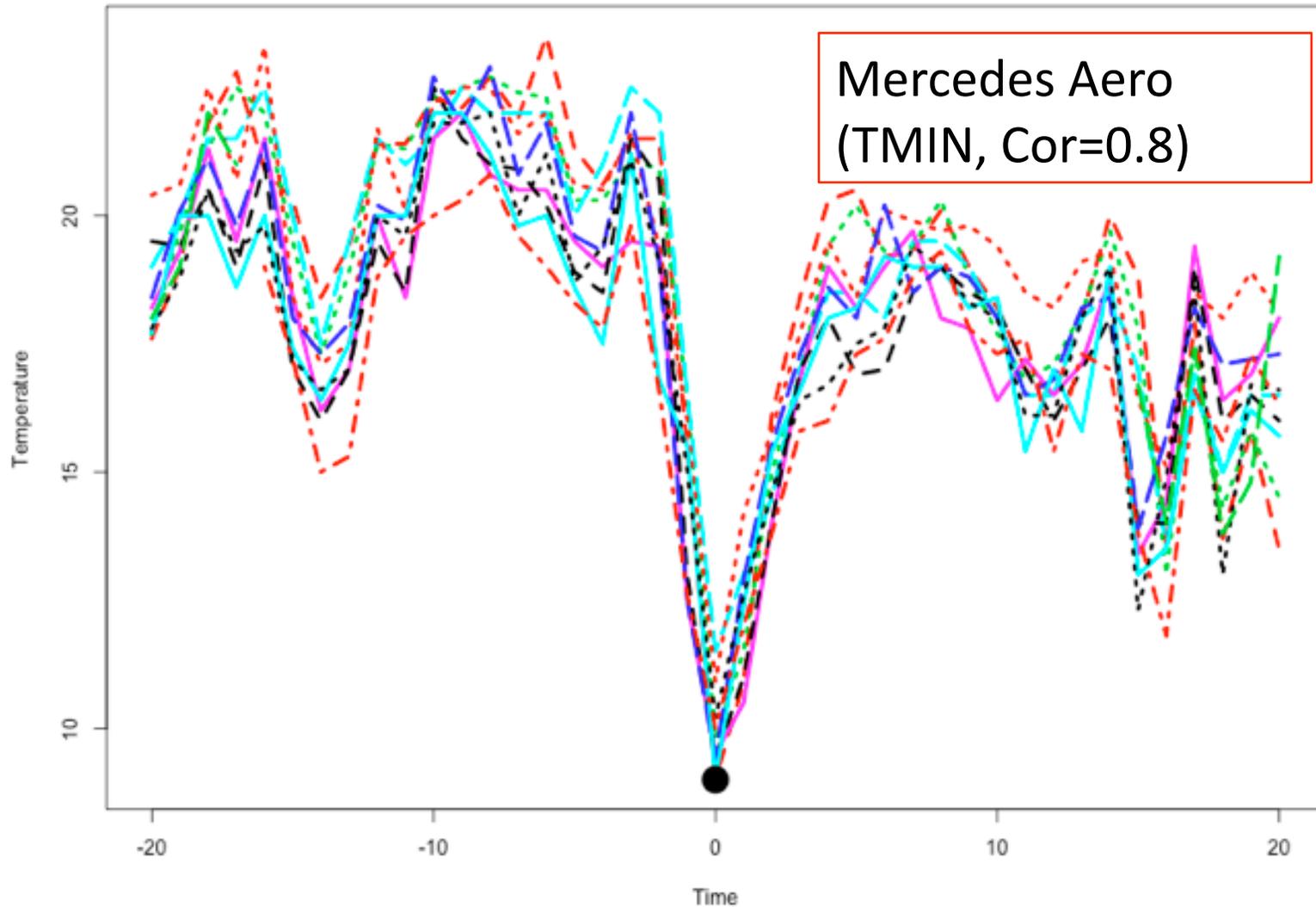
Huge Univariate Outlier

Station: 87171.TMAX. Date: 1960-08-02 Dist: 115.75 Contr: 1 #Stations: 2 #Observations: 41



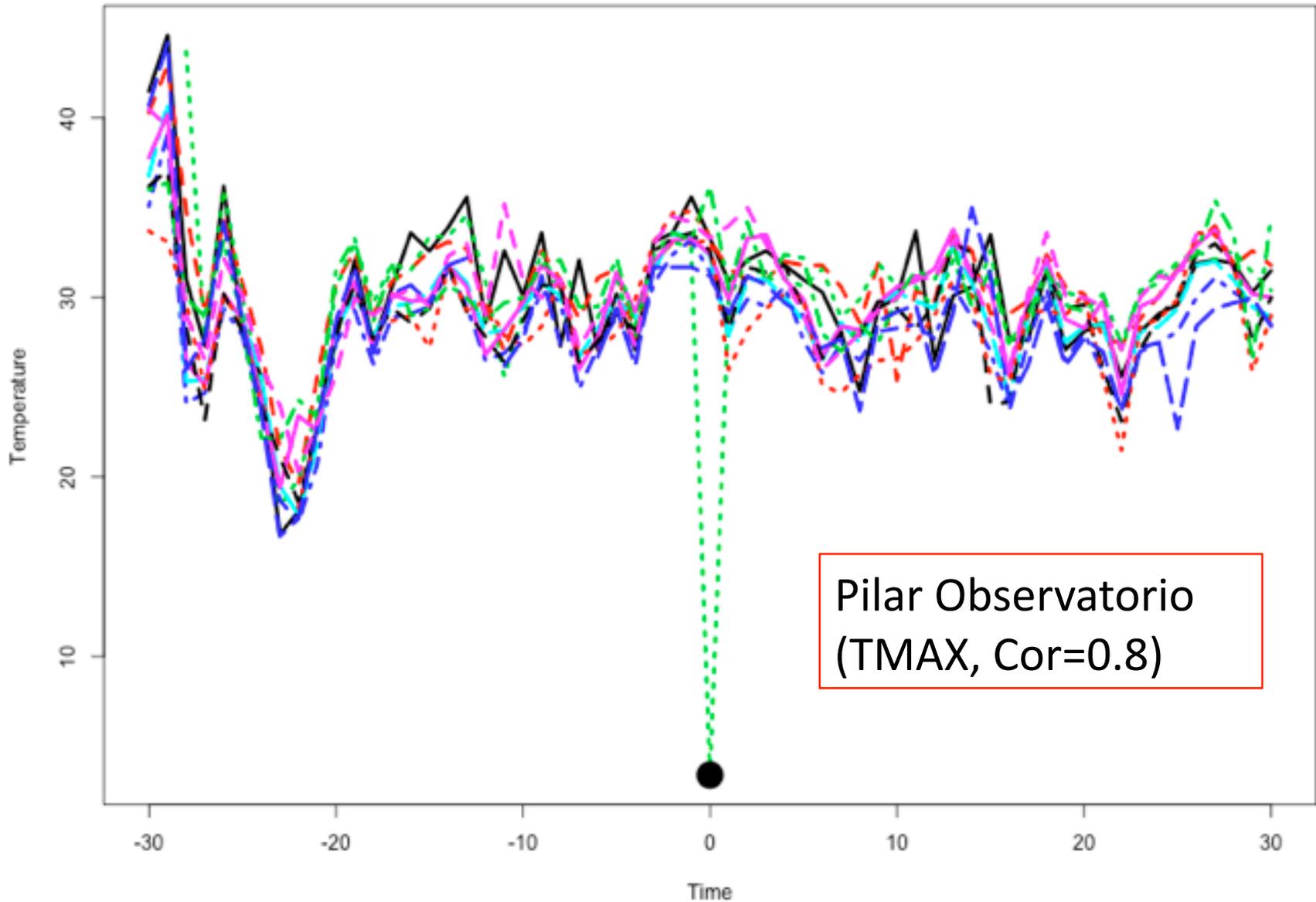
Multivariate Outlier: Odd Vs. Wrong

Station: 87158.TMIN. Date: 1974-03-17 Dist: 30.25 Contr: 0.24 #Stations: 10 #Observations: 37

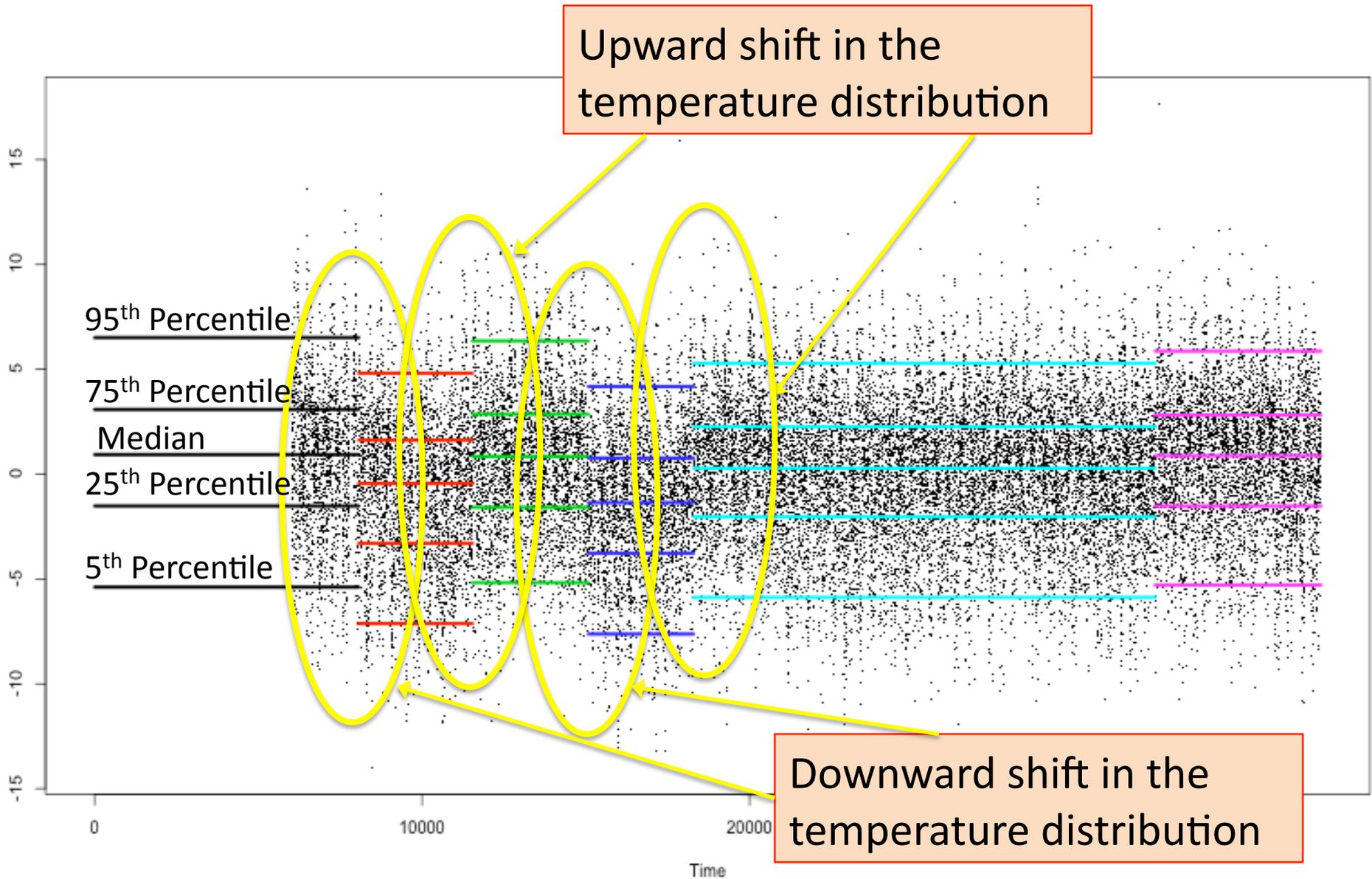


Zero Instead of Missing Value

Station: 87335.TMAX. Date: 1963-02-02 Dist: 143.71 Contr: 0.99 #Stations: 11 #Observations: 51



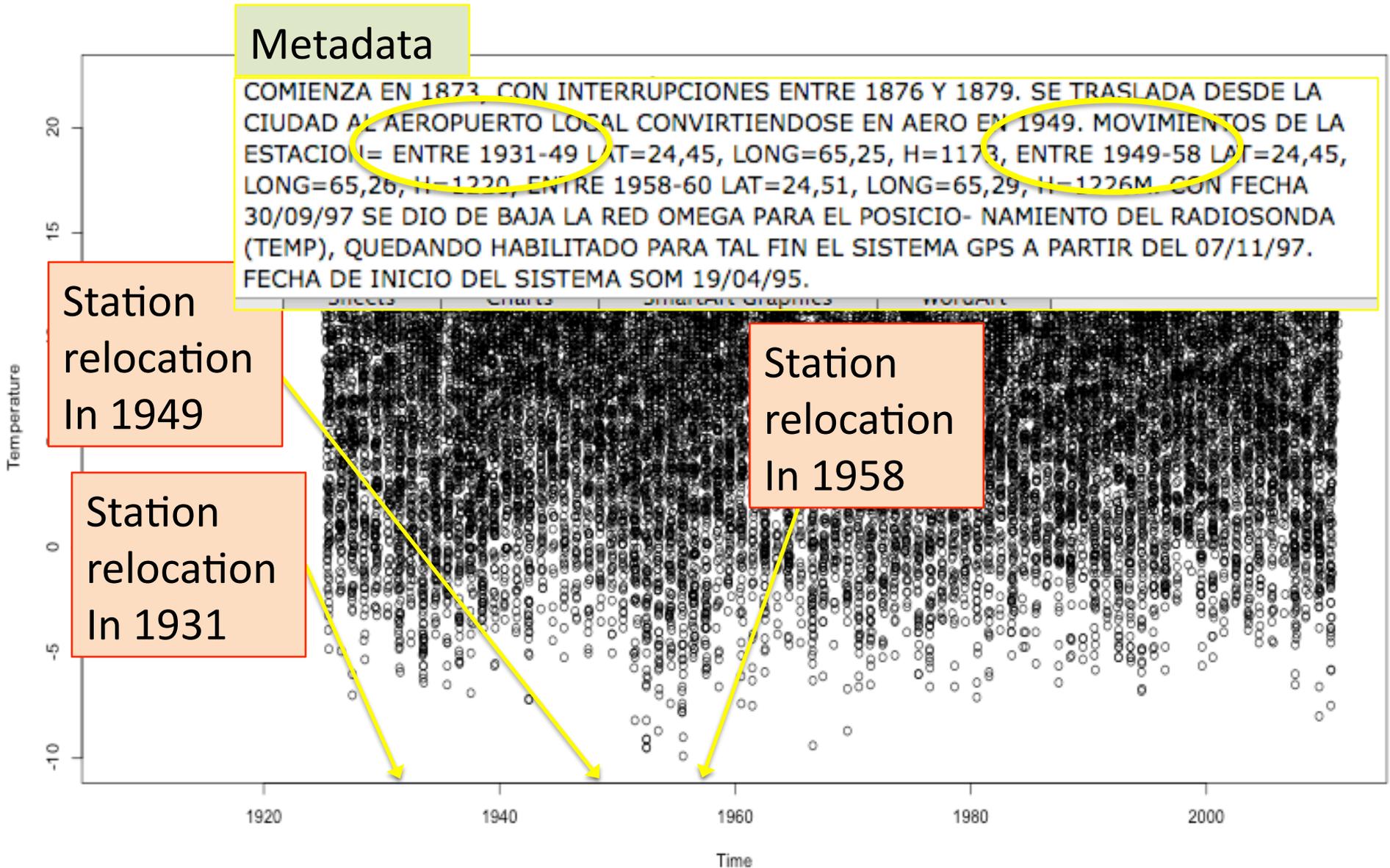
What is an Inhomogeneity



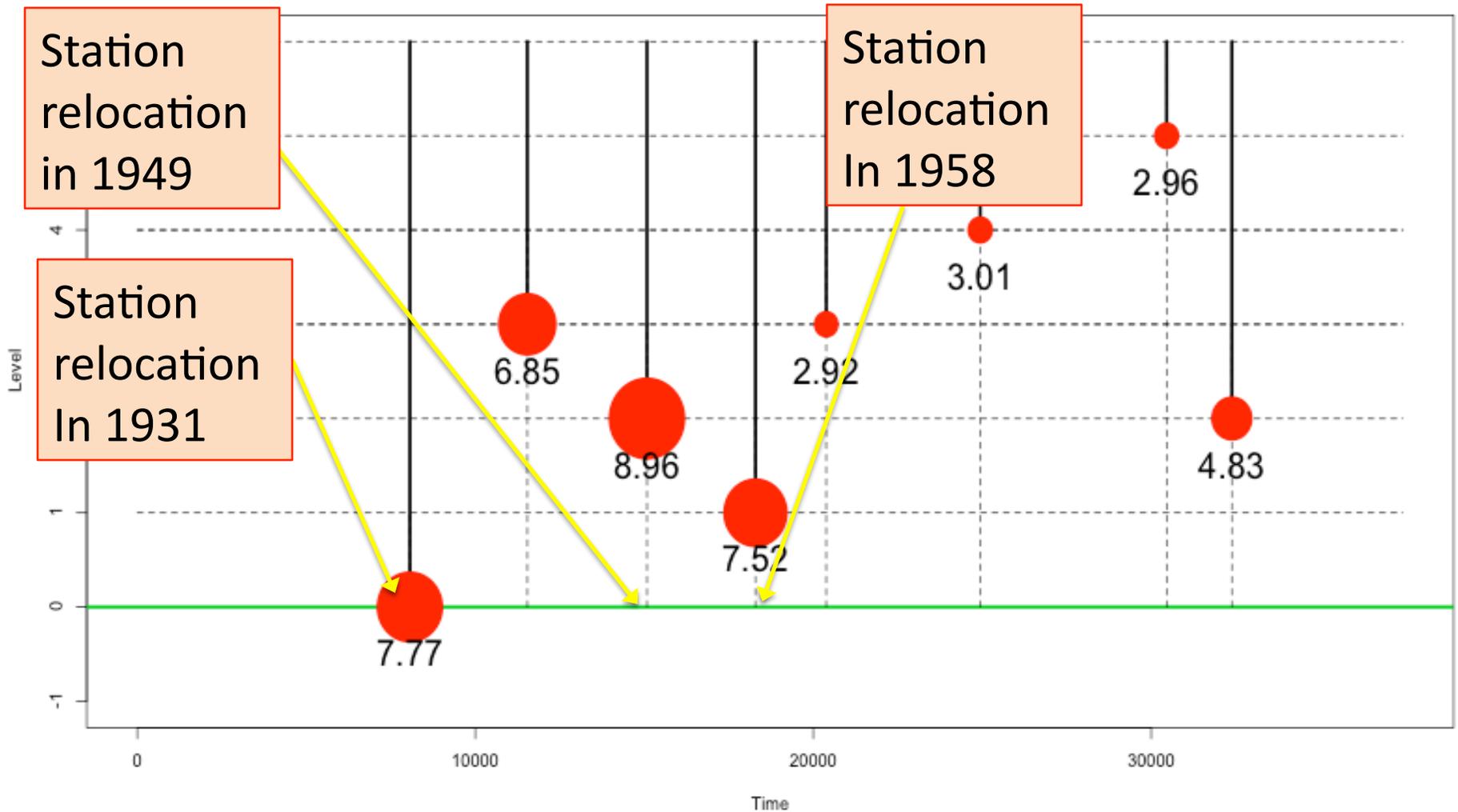
Inhomogeneties Detection Method

- Still under development
- Capable of:
 - Working with multiple stations simultaneously
 - 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)

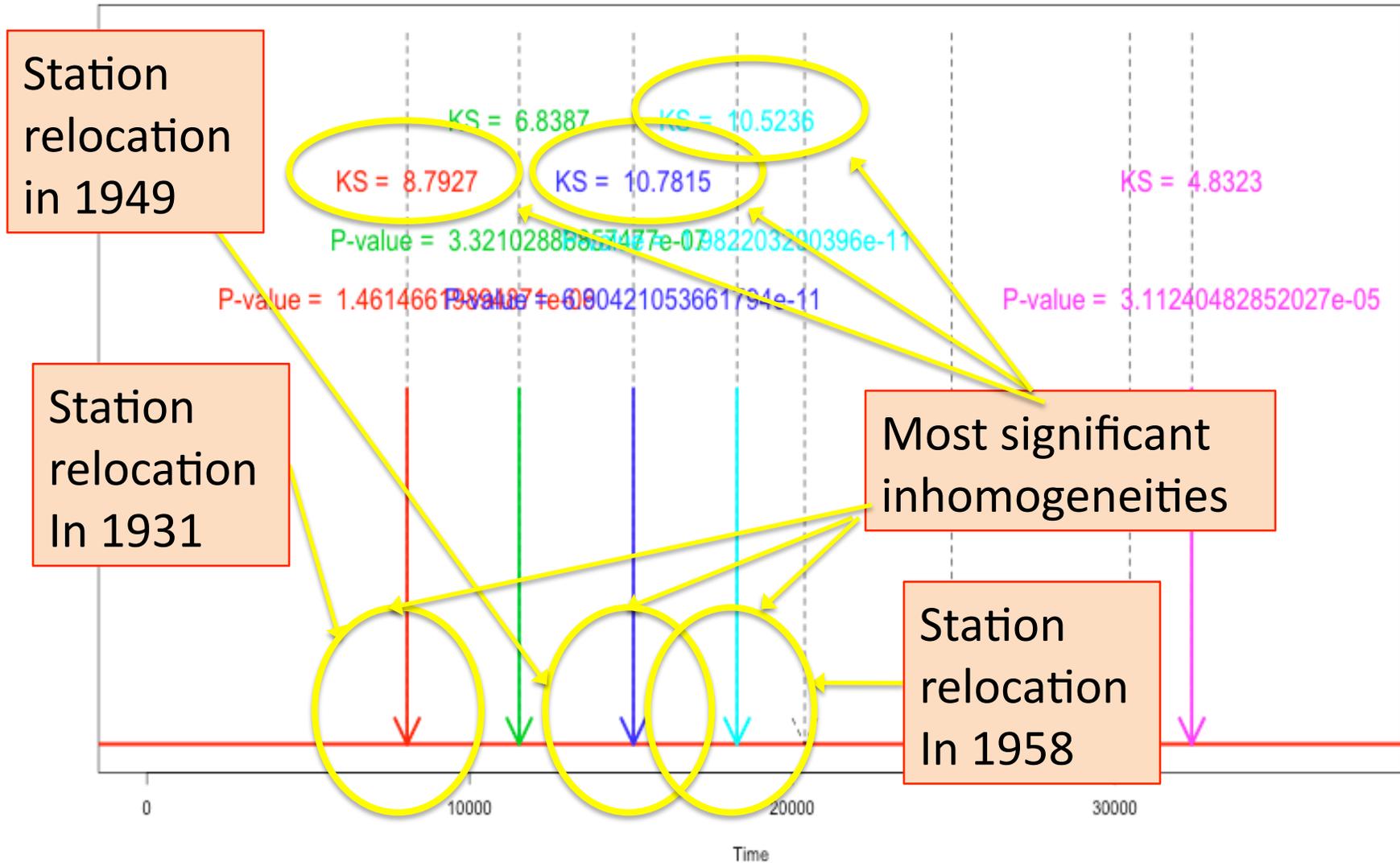


Growing the time partition tree



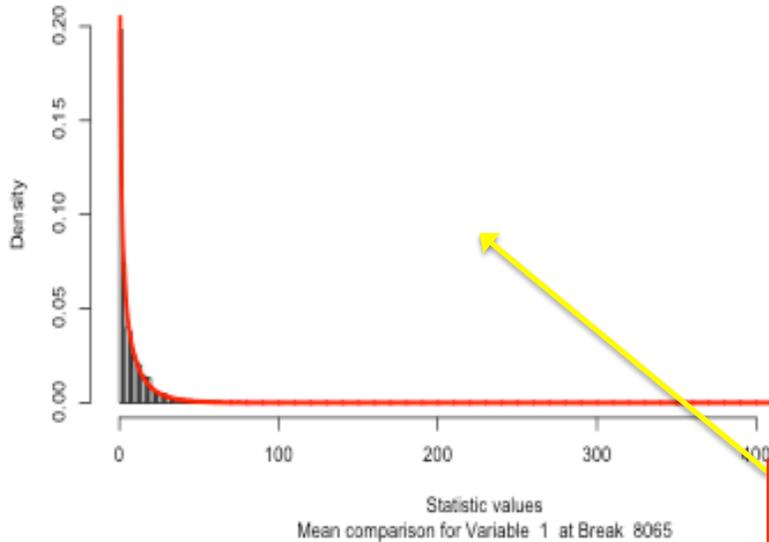
Significant breakpoints detection

Breakpoints detected

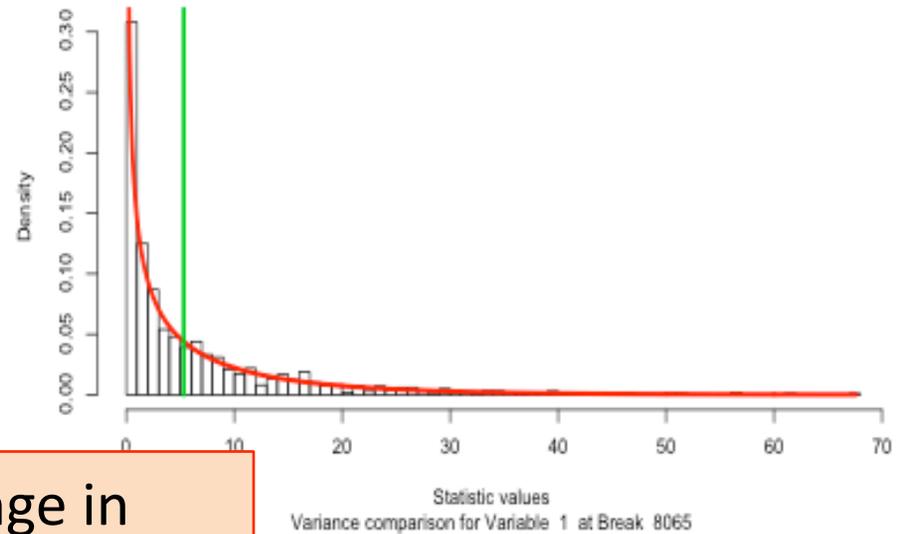


Determining the kind of inhomogeneity

Comparison for means, P-value = 0

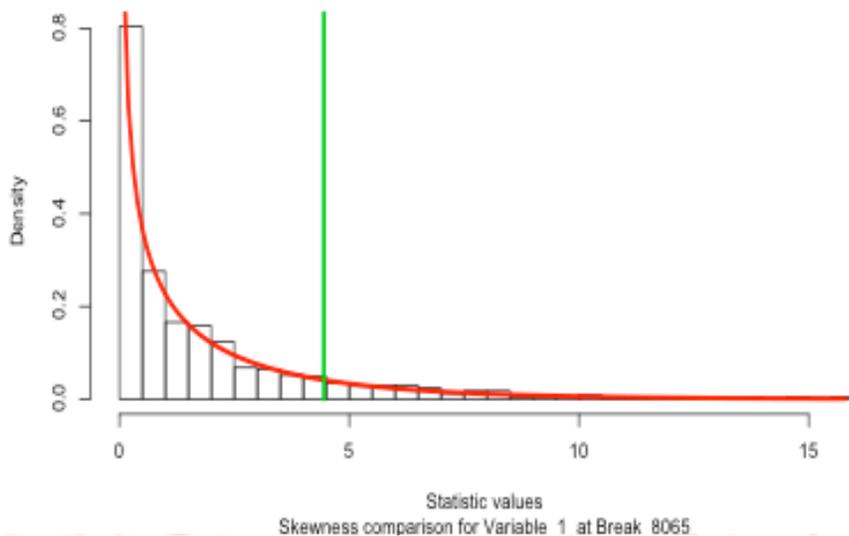


Comparison for variances, P-value = 0.356811

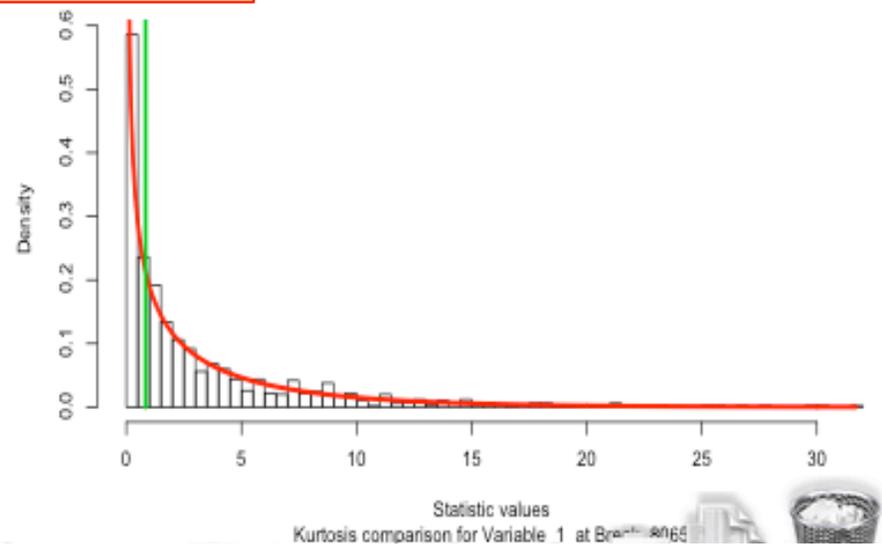


Change in mean in 1931

Comparison for skewness, P-value = 0.11582

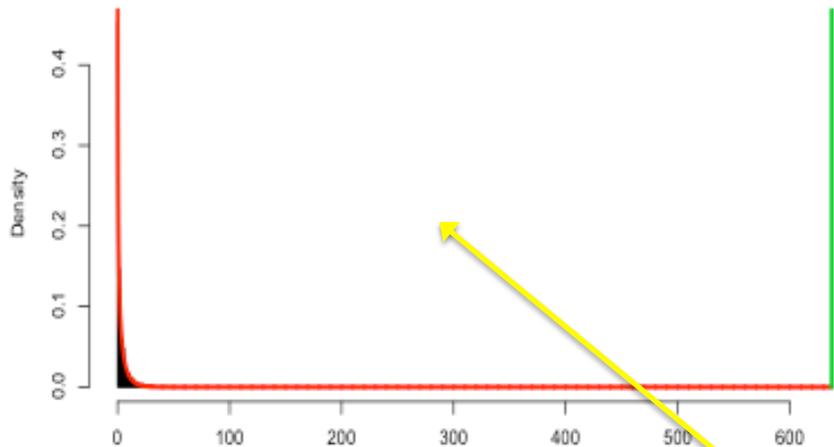


Comparison for kurtosis, P-value = 0.616774



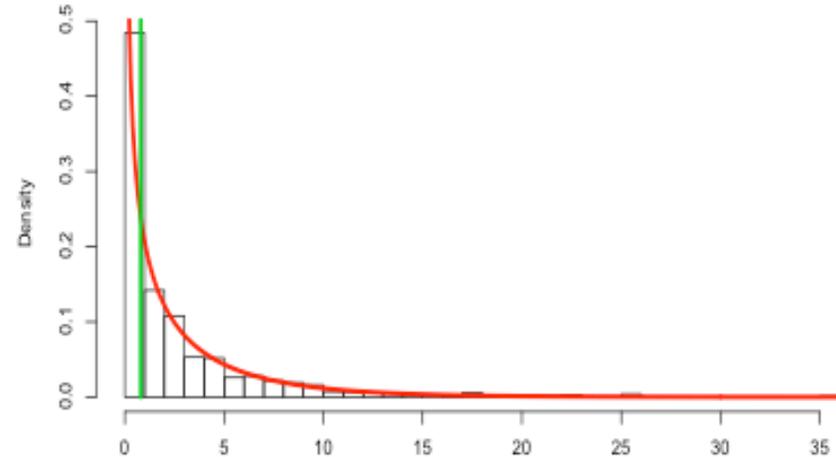
Determining the kind of inhomogeneity

Comparison for means, P-value = 0



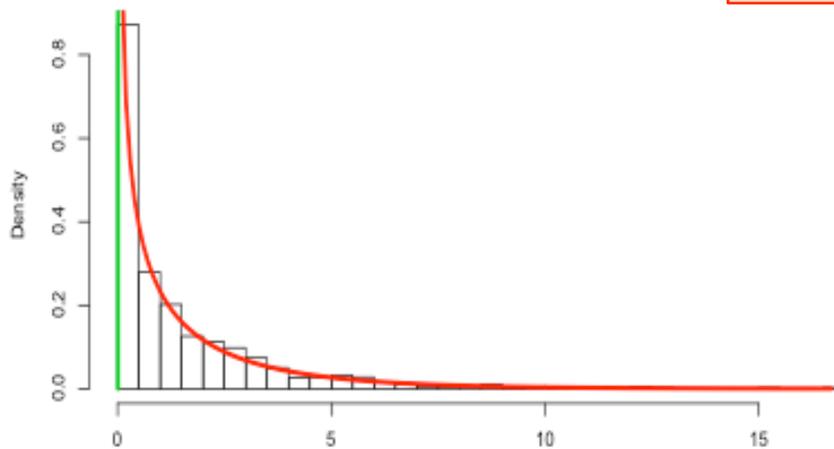
Statistic values
Mean comparison for Variable 1 at Break 15080

Comparison for variances, P-value = 0.589231



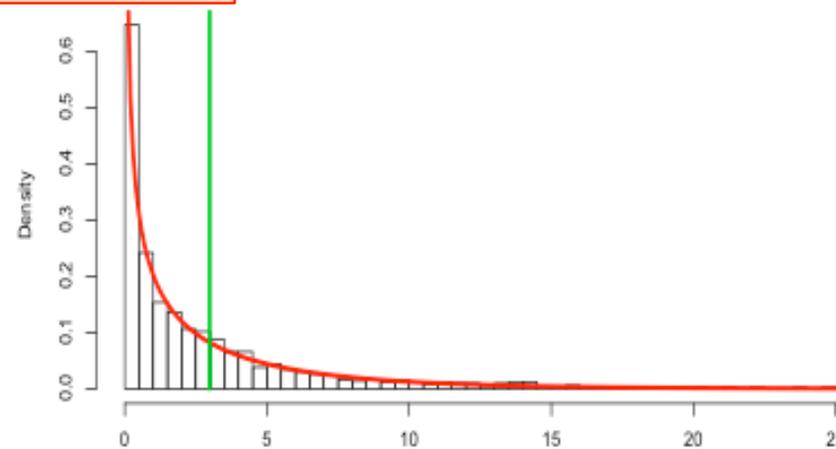
Statistic values
Variance comparison for Variable 1 at Break 15080

Comparison for skewness, P-value = 0.924964



Statistic values
Skewness comparison for Variable 1 at Break 15080

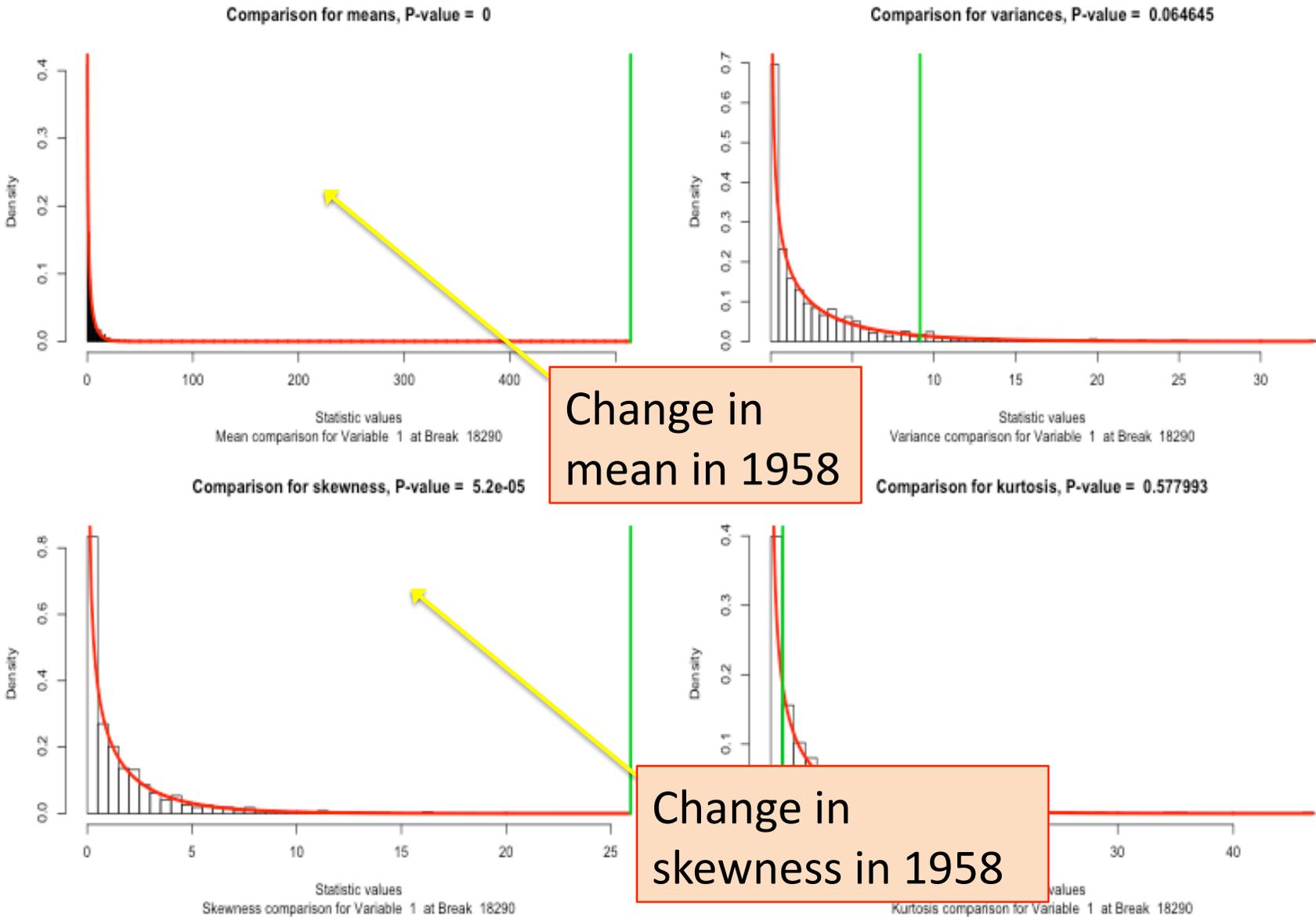
Comparison for kurtosis, P-value = 0.301712



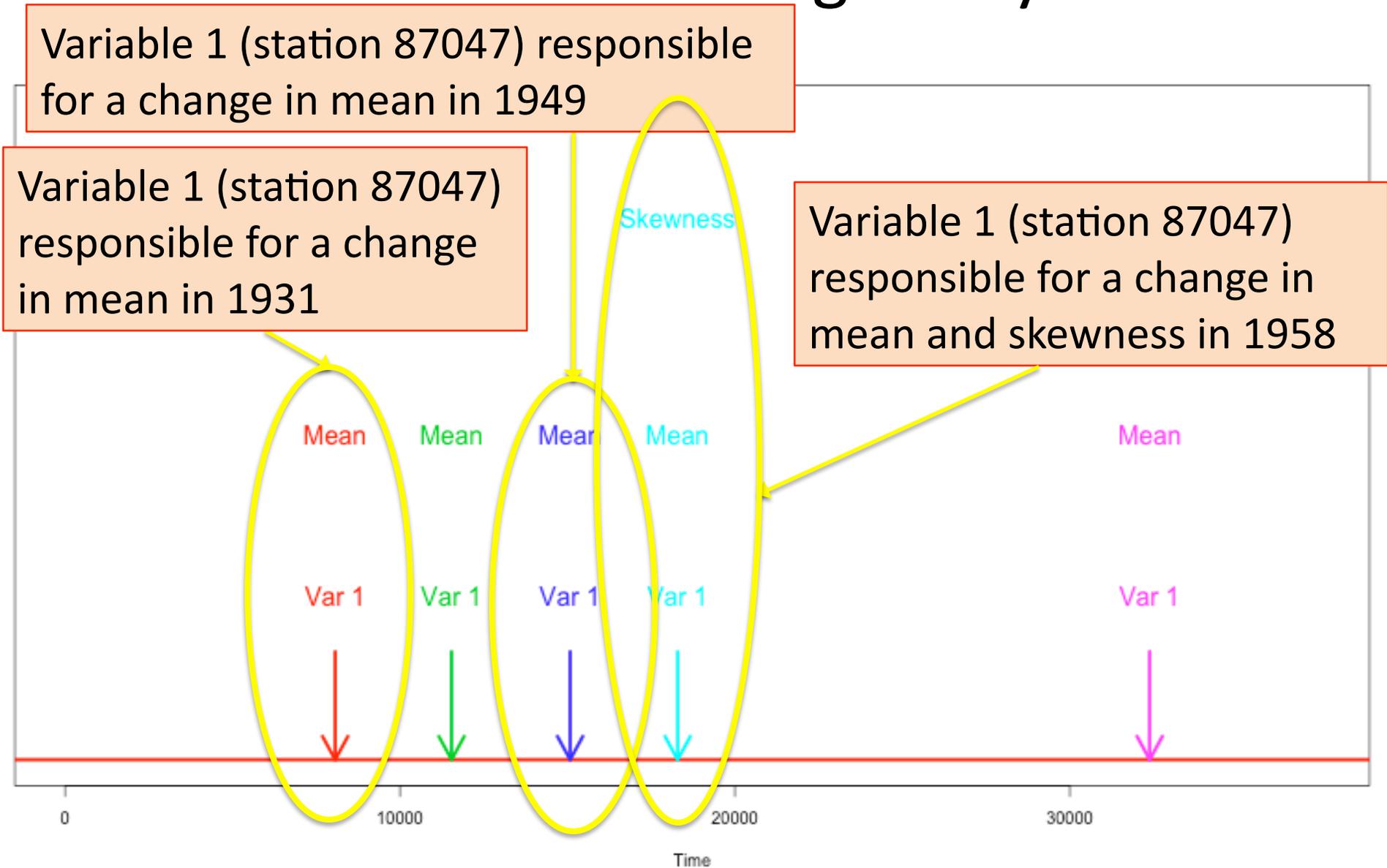
Statistic values
Kurtosis comparison for Variable 1 at Break 15080

Change in mean in 1949

Determining the kind of inhomogeneity



Identification of culprit station and kind of inhomogeneity



Showing the distributional changes

Inhomogeneities of variable 1

