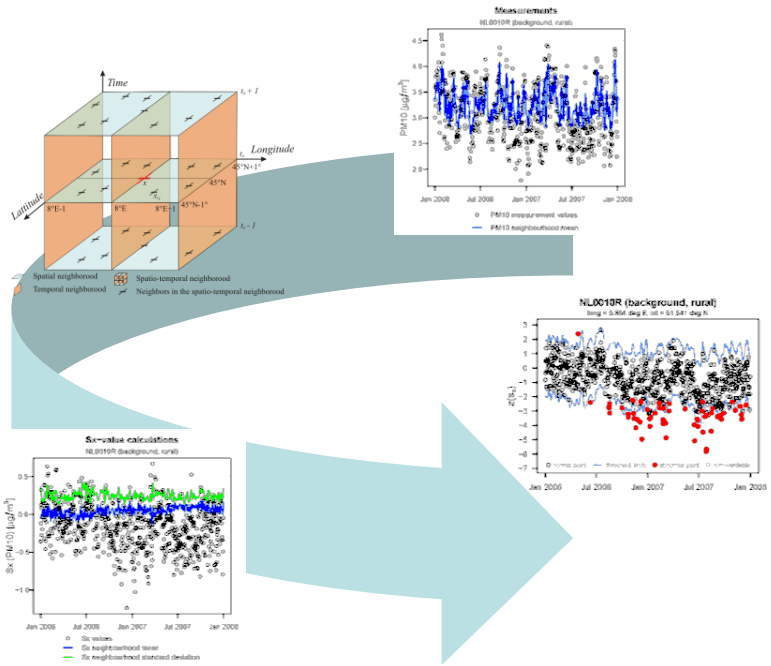
European Commission

Automated Screening of Spatio-Temporal Outliers in AirBase Records



Oliver Kracht and Michel Gerboles

European Commission - Joint Research Centre
I – 21026 Ispra (VA)
www.jrc.ec.europa.eu

FAIRMODE
Forum for air quality modelling in Europe

Fairmode Technical Meeting
24th and 25th June 2015

Aveiro - Portugal

Joint Research Centre

Intro

Objectives:

- Present a screening tool for abnormal values of ambient air quality monitoring stations
- How to use these results?

Methodology:

- "Smooth Spatial Attribute Method" (S(x)-outlier)

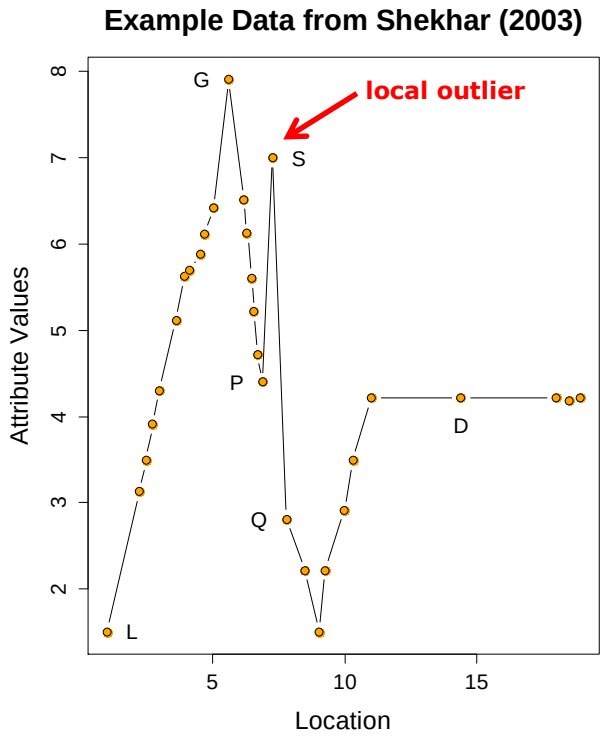
  (first developed for traffic sensors by Lu et al 2003 & Shekhar et al 2003)

Applications:

- Screening AirBase records of daily PM_{10} and NO_2 values

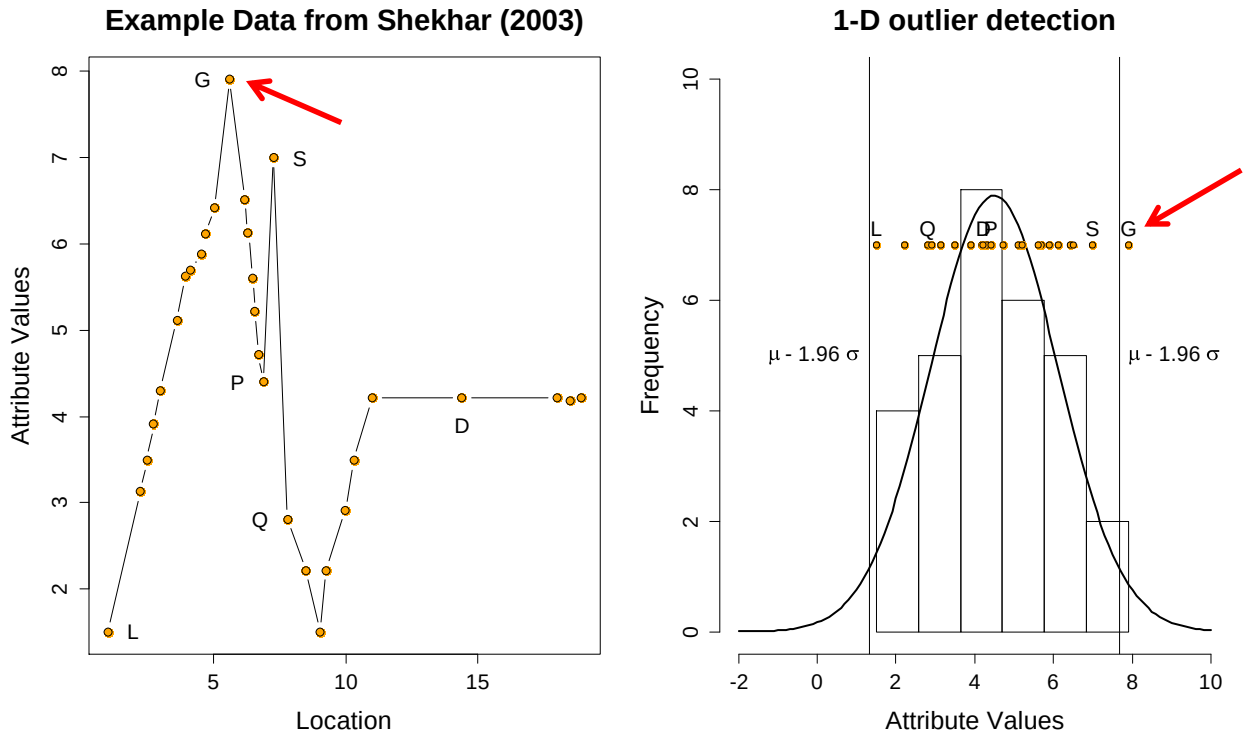
A bit of taxonomy for spatio-temporal outliers

- What is a local outlier?



A bit of taxonomy for spatio-temporal outliers

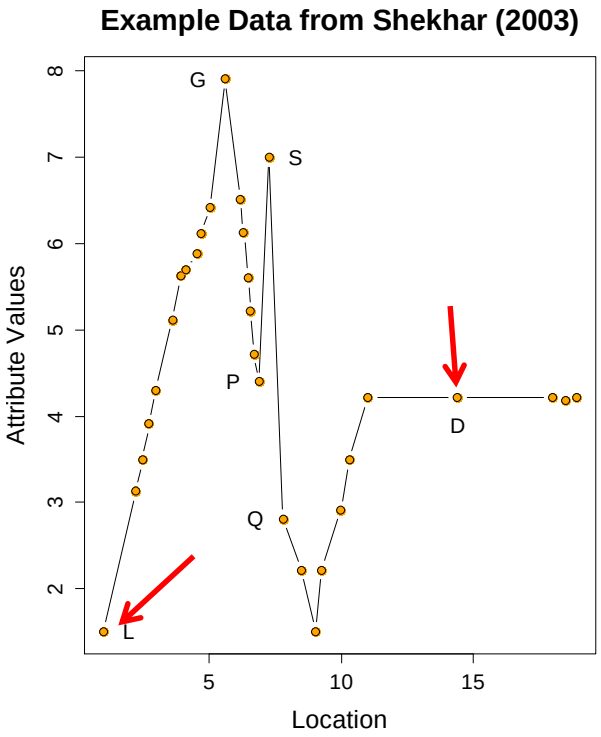
- What is a local outlier?

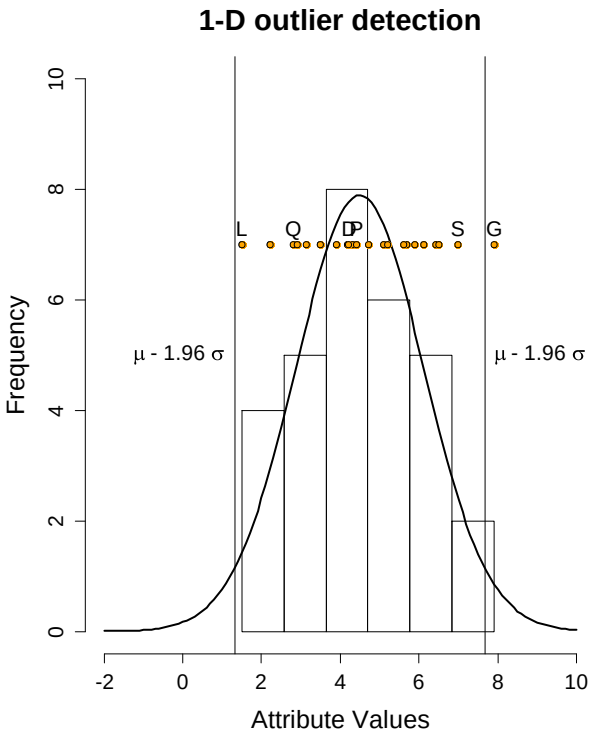


The outlier detected using a 1-D approach is point G (being **globally extreme**).

A bit of taxonomy …

- What is a local outlier?

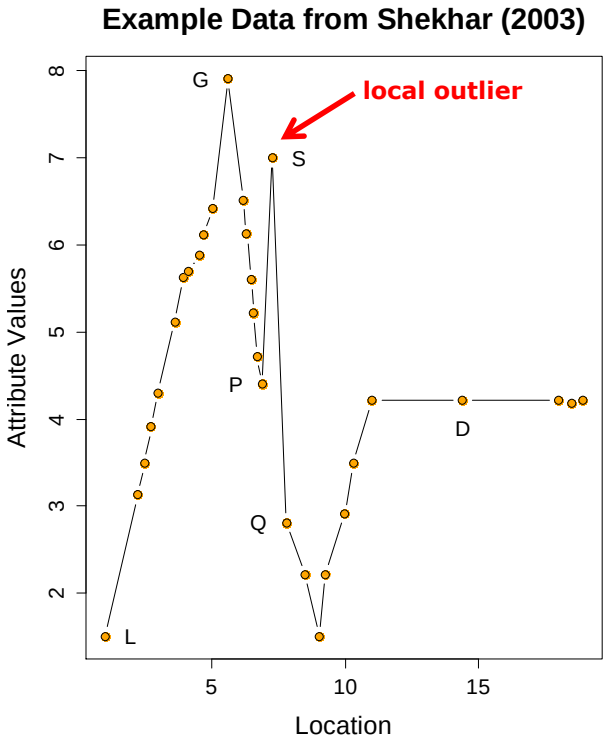


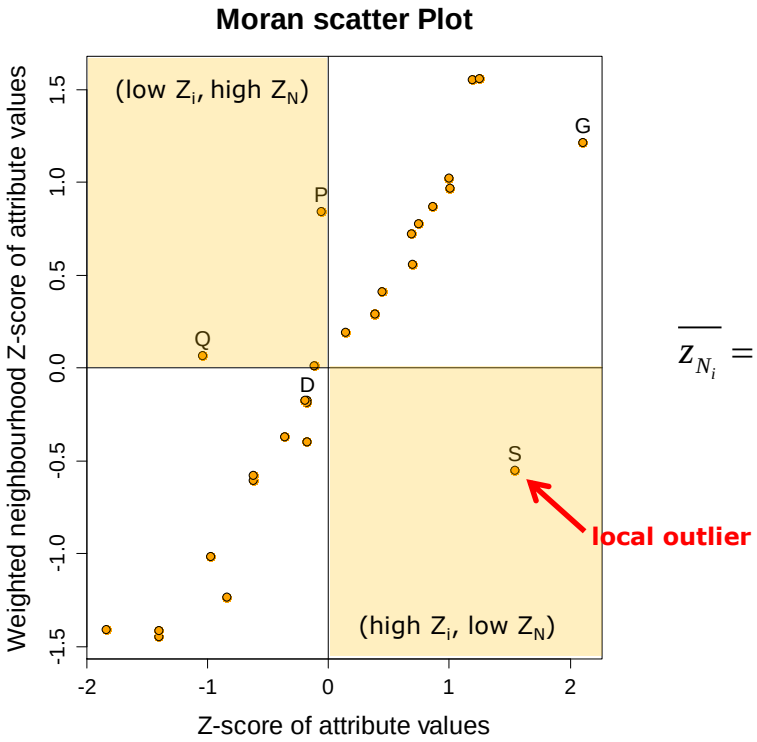


Multi dimensional distance or density based approaches might detect points L and D (by **degree of isolation**).

A bit of taxonomy …

- Bi-partite multi dimensional tests are separating spatial attributes from non-spatial attributes (**spatial outliers**)

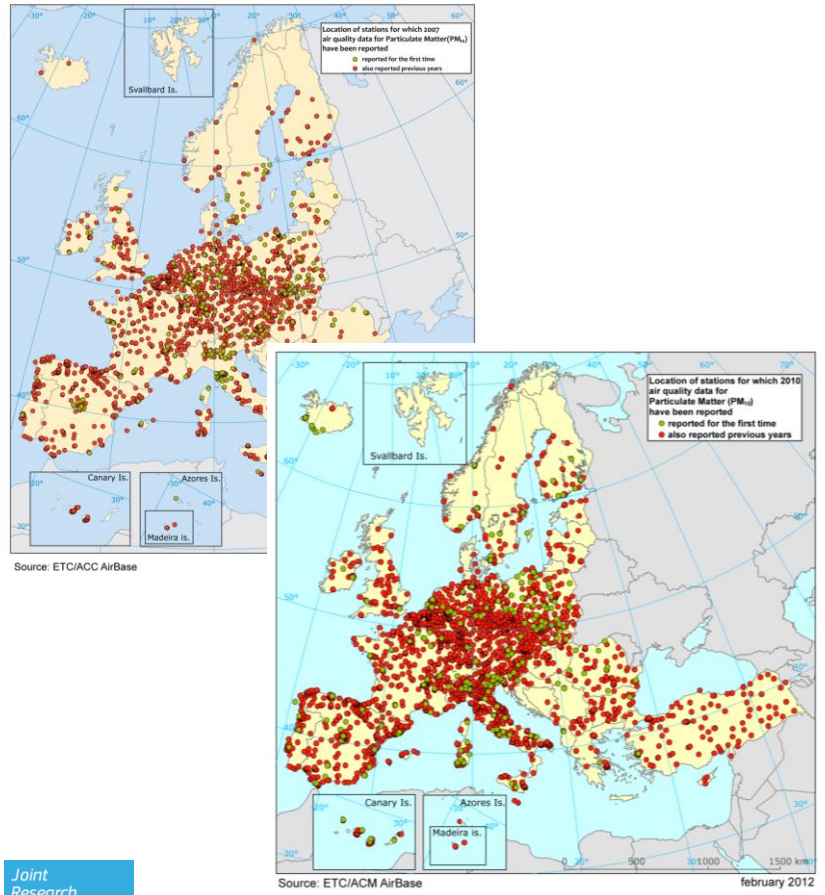




$$\overline{z_{N_i}} = \frac{\sum_{N_i} w_i \cdot z_i}{\sum_{N_i} w_i}$$

$$z_i = \frac{x_i - \bar{x}}{std(x)}$$

How to treat spatially distributed time series e.g. collected in AirBase?



Adaption of the Smooth Spatial Attribute Method

- Proposed for traffic sensors by Lu et al. 2003 & Shekhar et al. 2003

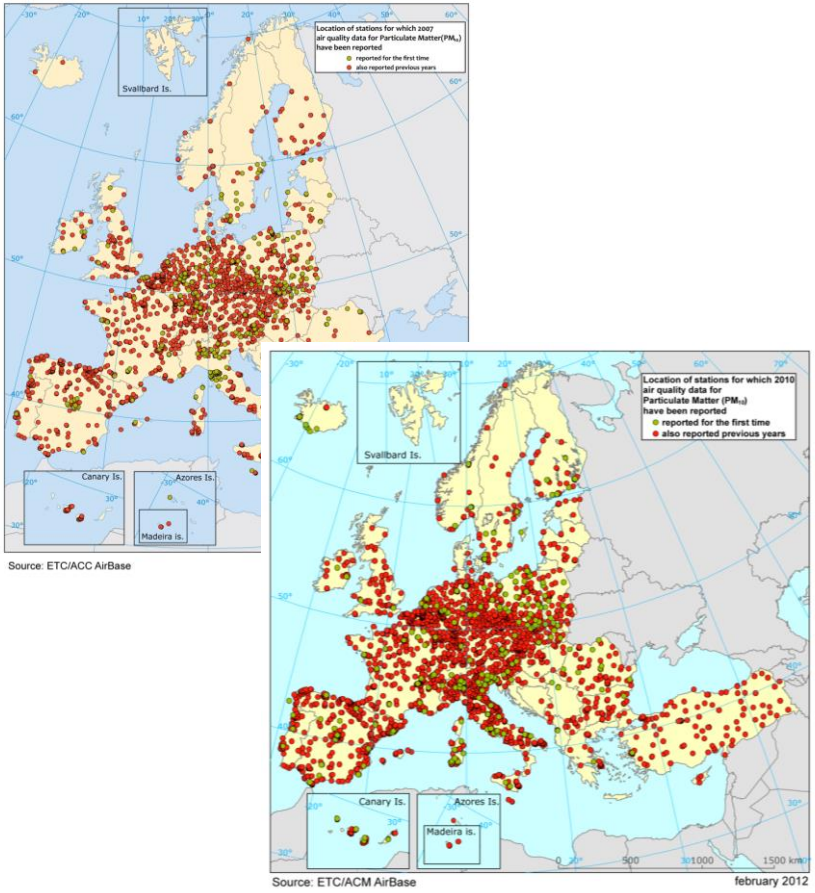
  Lu, CH.-T., D. Chen & Y. Kou, 2003: Detecting Spatial Outliers with Multiple Attributes. ICTAI'03, IEEE 2003.

  Shekhar, S., CH.-T. Lu & P. Zhang, 2003: A Unified Approach to Detecting Spatial Outliers. GeoInformatica, 7(2), 139-166.

- 1st quantify how the measurement value of a station deviates from the corresponding values observed within its spatio-temporal neighbourhood ("S(x)-value")

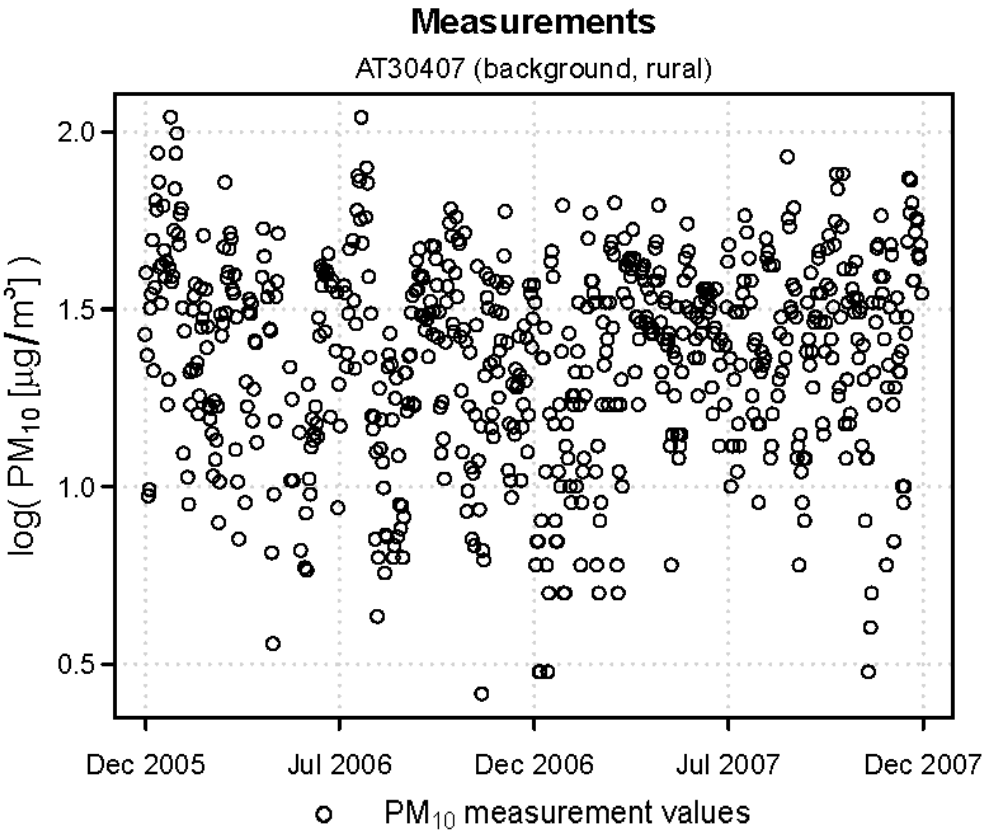
- 2nd compare this S(x)-deviation to the corresponding S(x)-deviations observed for the station's neighbours

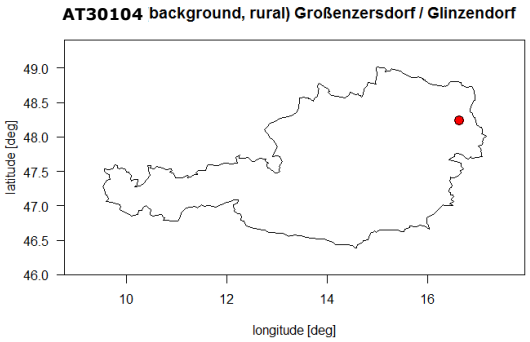
Focus of this Exercise:



- 2001 – 2010 records from AirBase v.8
- 18 selected country sets
- daily PM_{10} and NO_2 values
- station type "Background"
- all area types (urban, suburban and rural)

"In a nutshell" example for spatio-temporal outlier screening:

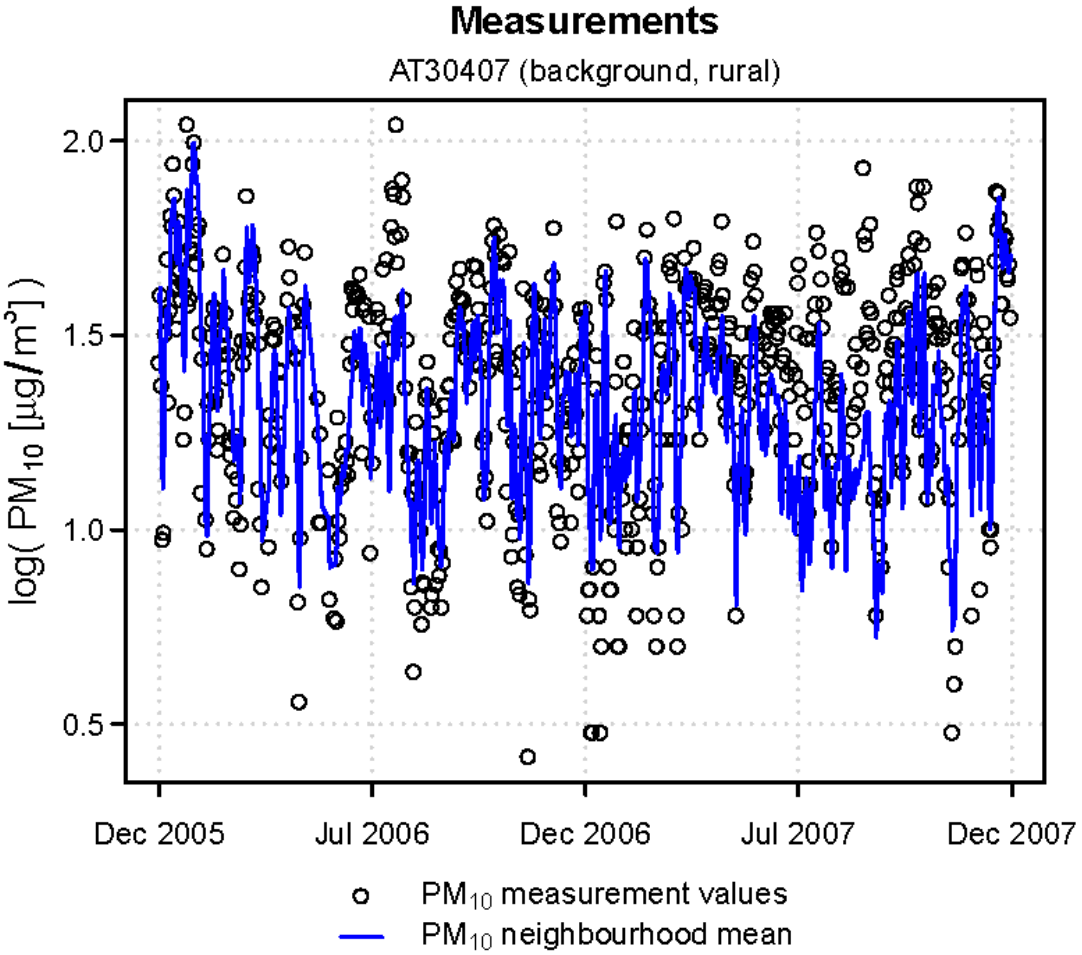


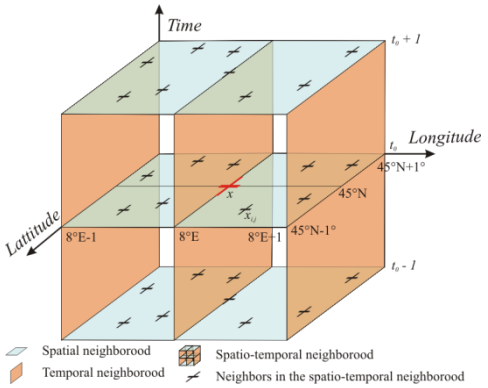


- ***Step 1:***

  $\log_{10}$ transformation of non-Gaussian data

Methods Example



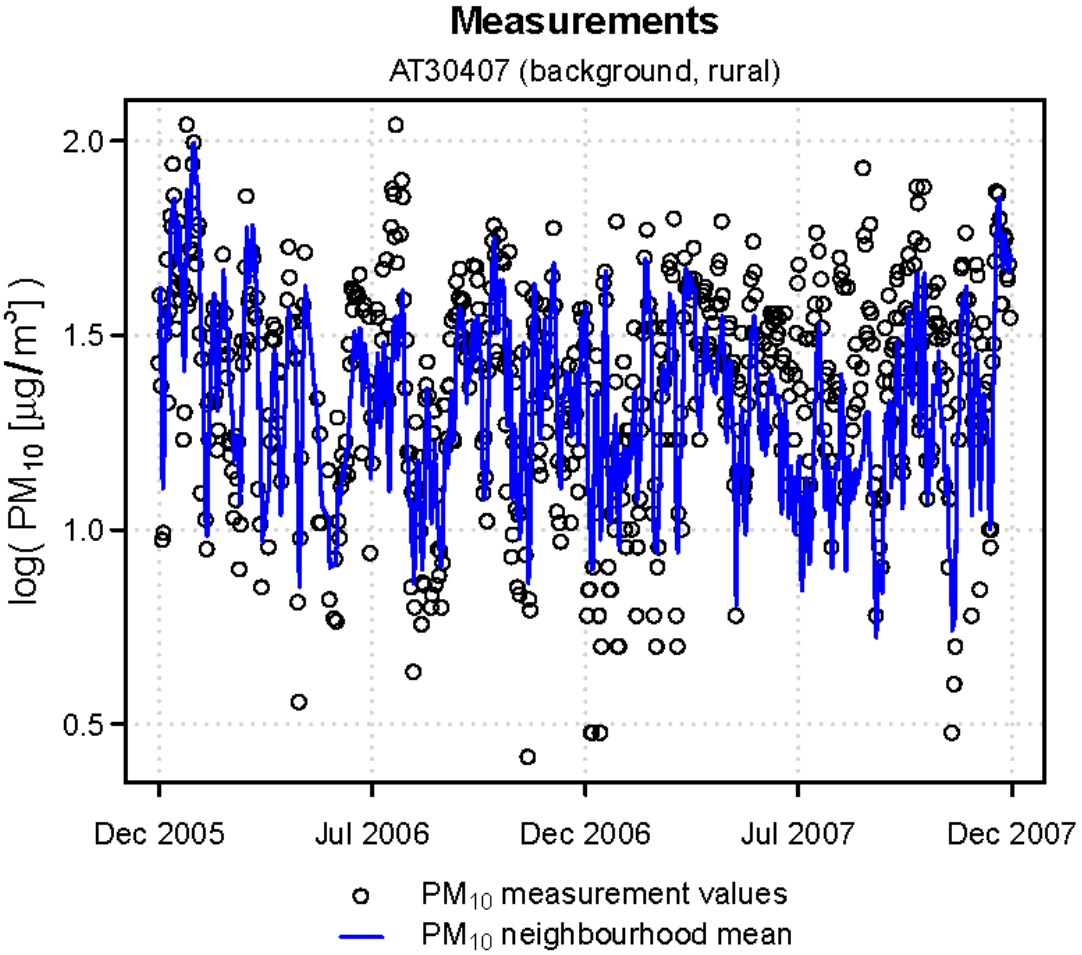


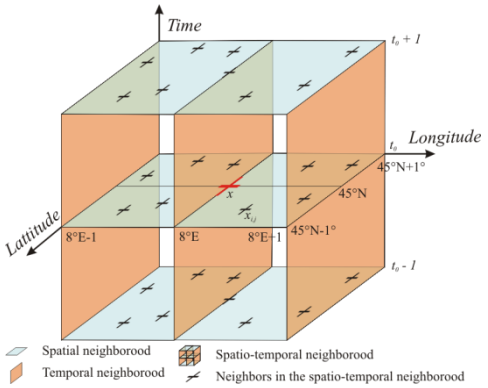
- ***Step 2:***

  Calculate neighbourhood mean.

  (weighted mean using the cubic of the inverse normalized Euclidian distance)

Methods Example





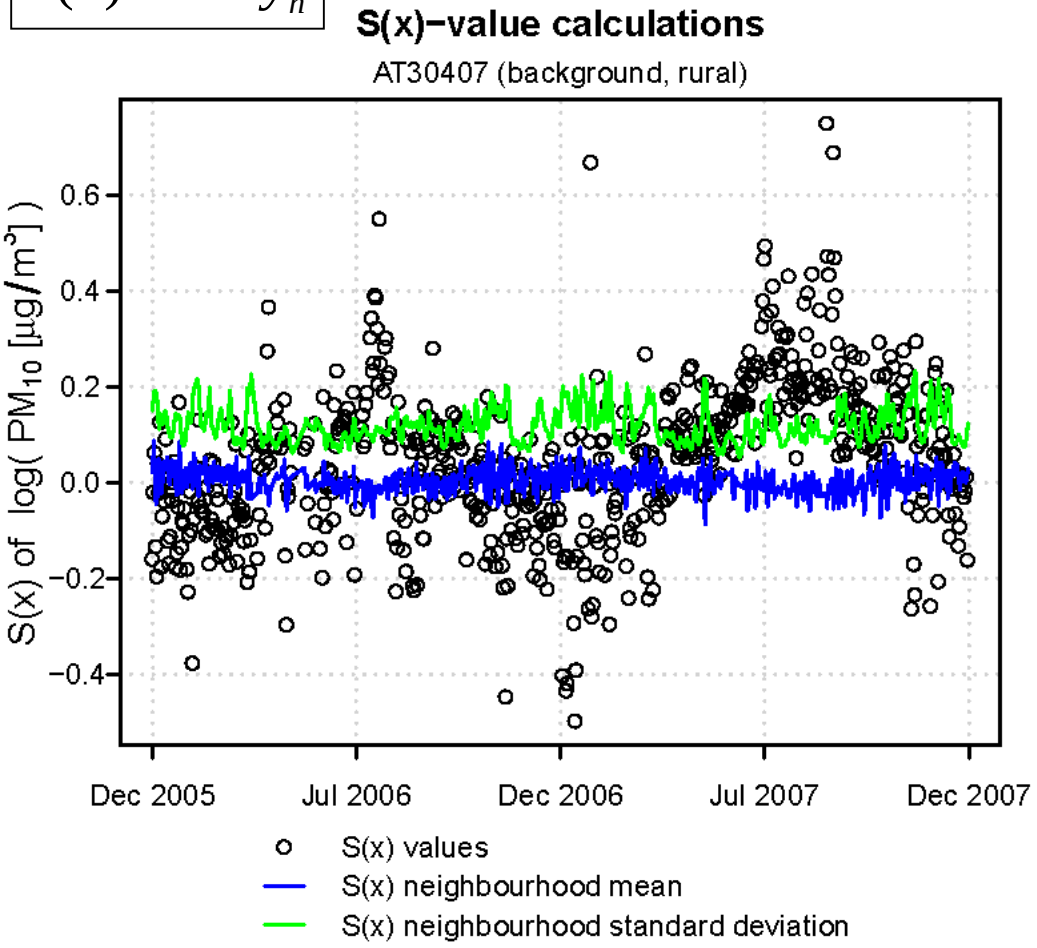
Step 3:

Calculate S(x) for all stations within their individual neighbourhoods.

$$S(x) = x - \overline{y_n}$$

Methods Example

$$S(x) = x - \overline{y_n}$$

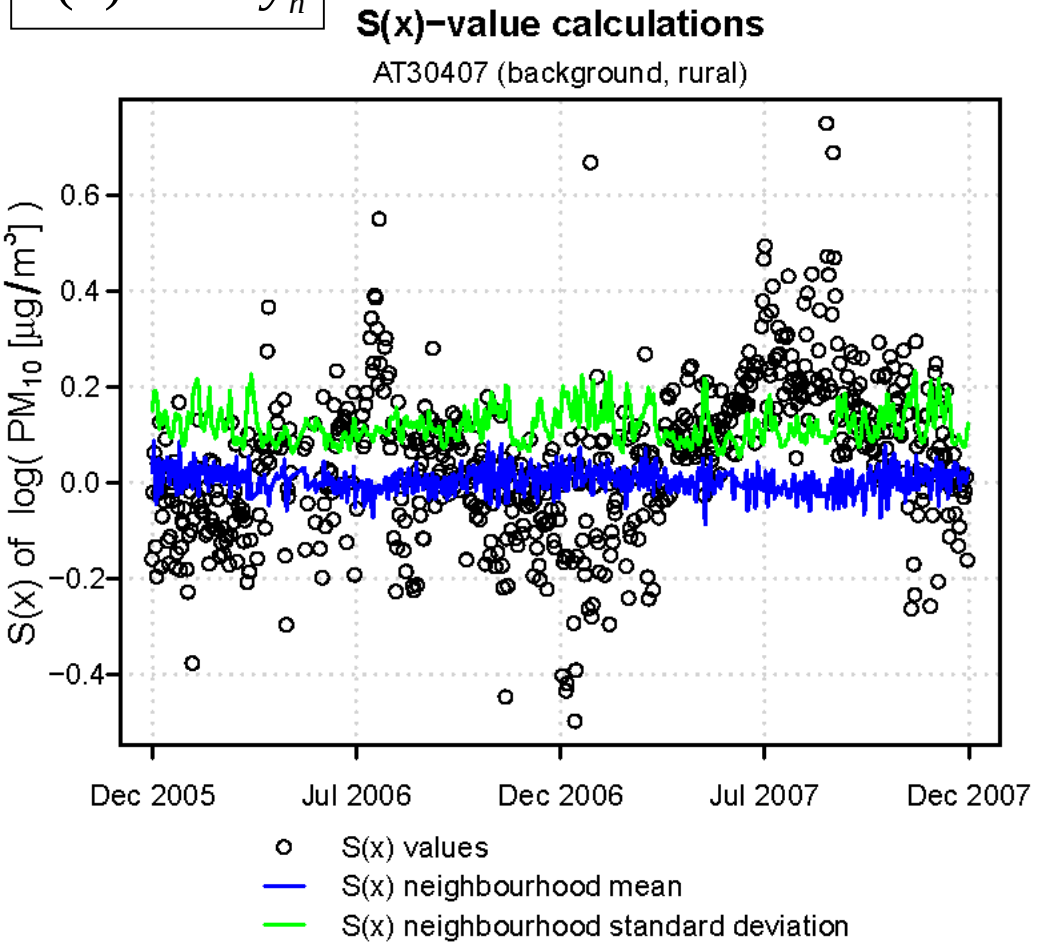


- ***Step 3:***

  Calculate S(x) for all stations within their individual neighbourhoods.

Methods Example

$$S(x) = x - \overline{y_n}$$



- ***Step 3:***

  Calculate S(x) for all stations within their individual neighbourhoods.

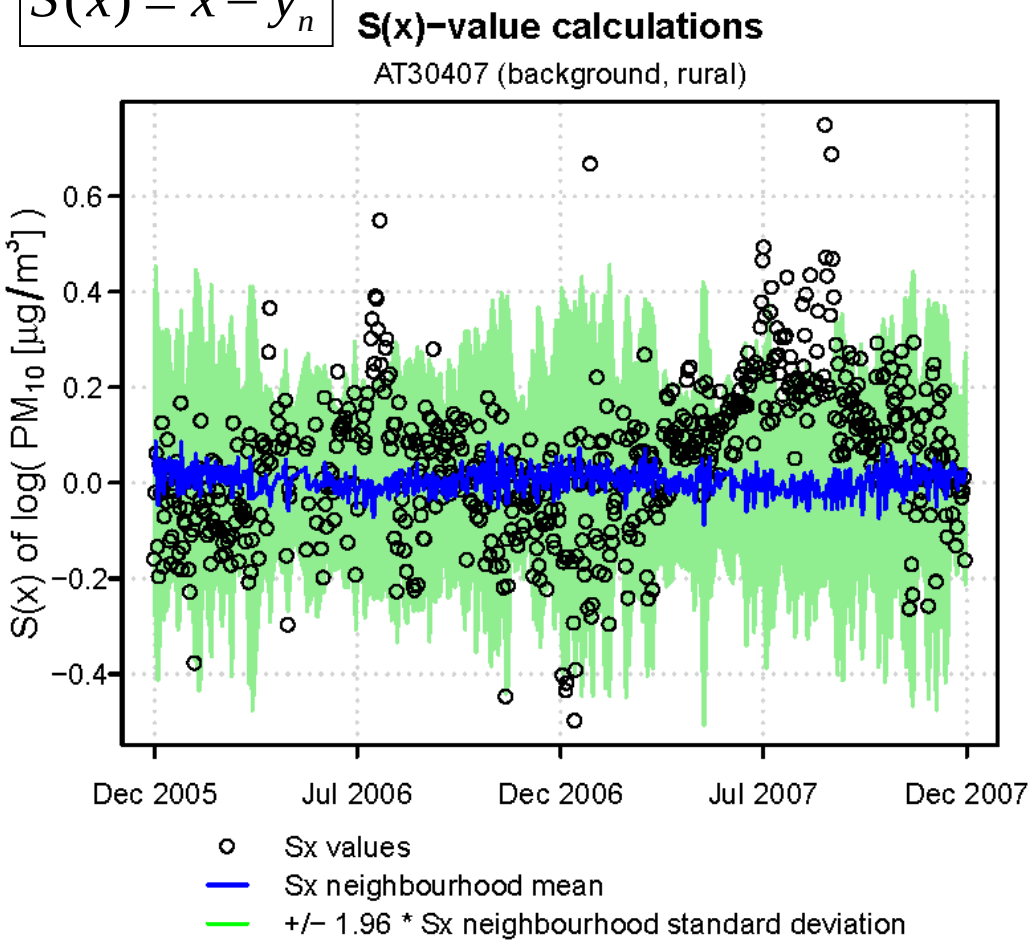
- ***Step 4:***

  For every neighbourhood, obtain the weighted mean and weighted standard deviation of *S(x)*

  - mean($S(x)_N$)
  - $s(S(x)_N)$

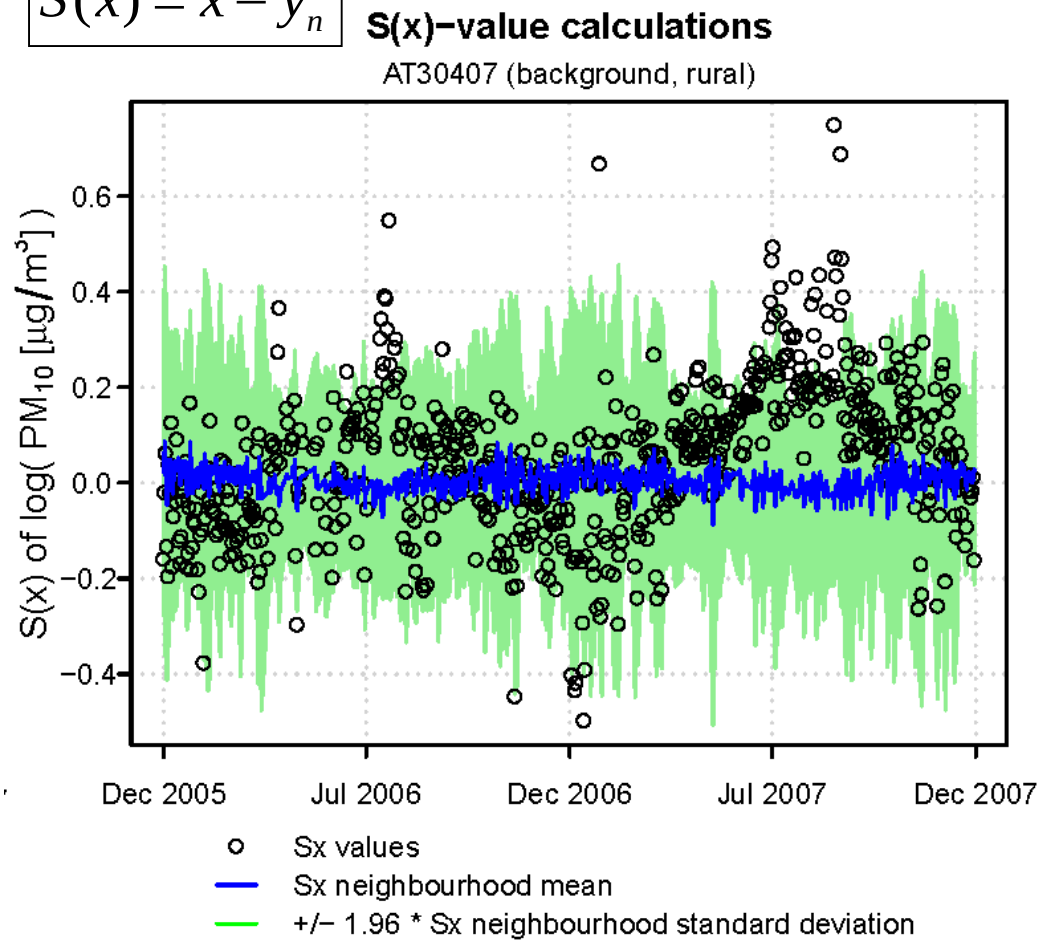
Methods Example

$$S(x) = x - \overline{y_n}$$



Methods Example

$$S(x) = x - \overline{y_n}$$



- ***Step 5:***

  Z-transform S(x) values of every station

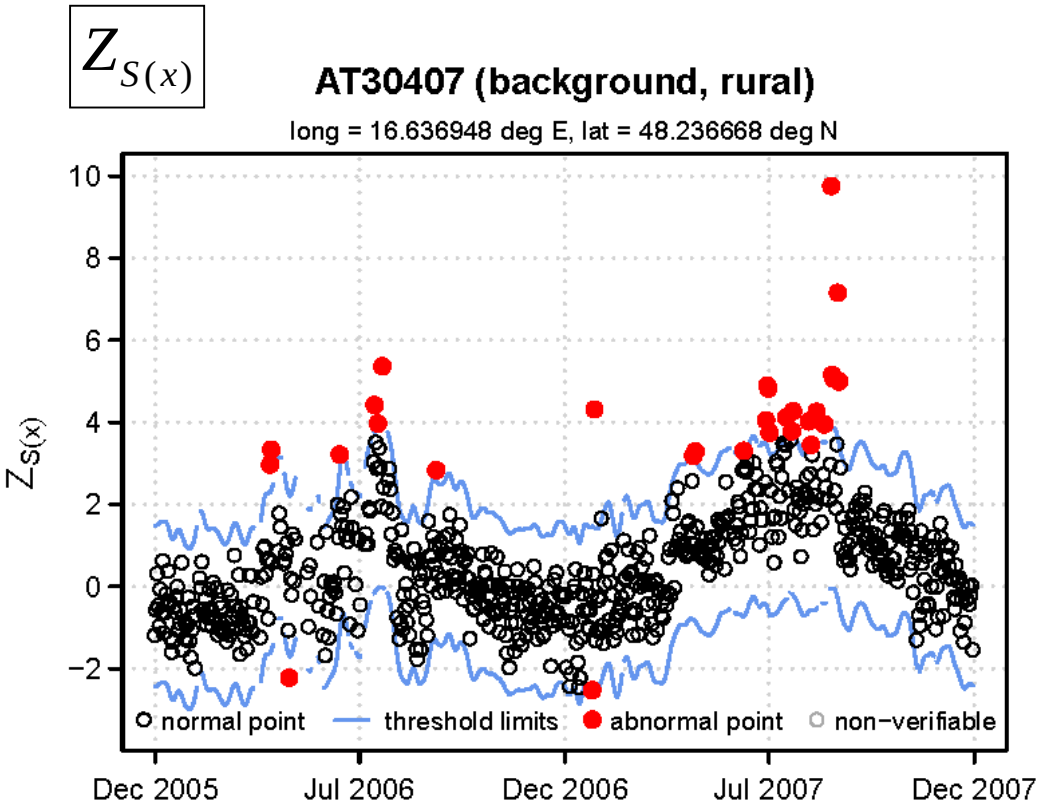
$$Z_{S(x)} = \frac{S(x) - \overline{S(x)_{N(x)}}}{s(S(x)_{N(x)})}$$

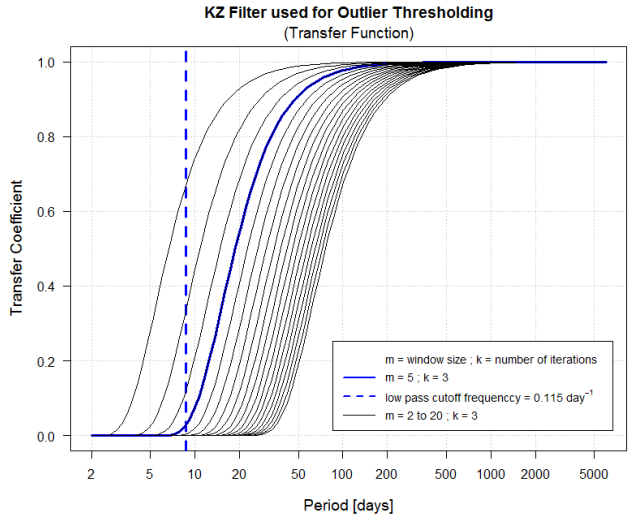
Methods Example

- **Step 6:**

Define a reference basis θ_{ref} and threshold limits.

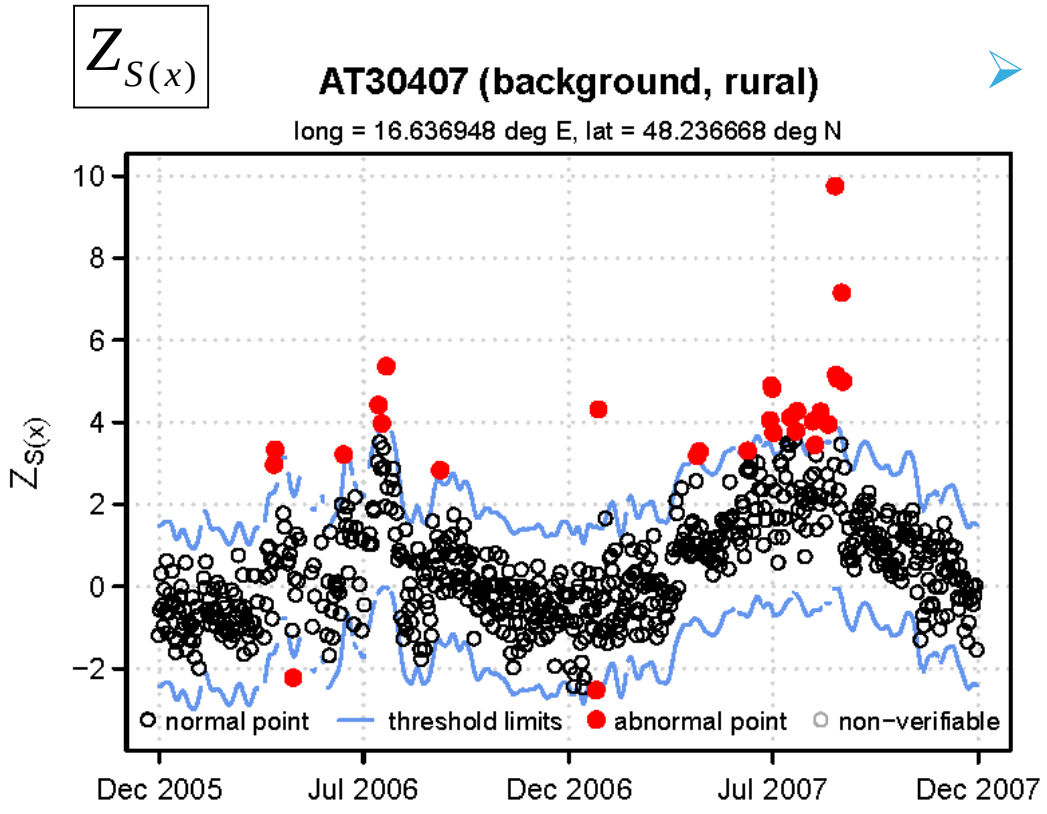
$$\theta = \theta_{ref} \pm \theta_{conf} = KZ(Z_{S(x)}) \pm \theta_{conf}$$





$KZ_{(m=5,\ k=3)}$ effectively removes signal components with a periodicity of less than ca 8.7 days

Methods Example



- ***Step 7:***

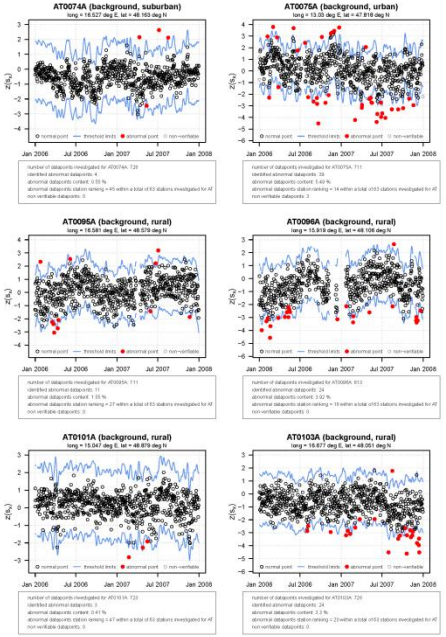
Test statistics for abnormal values searches for z_i values exceeding the upper/lower limits chosen as a reference.

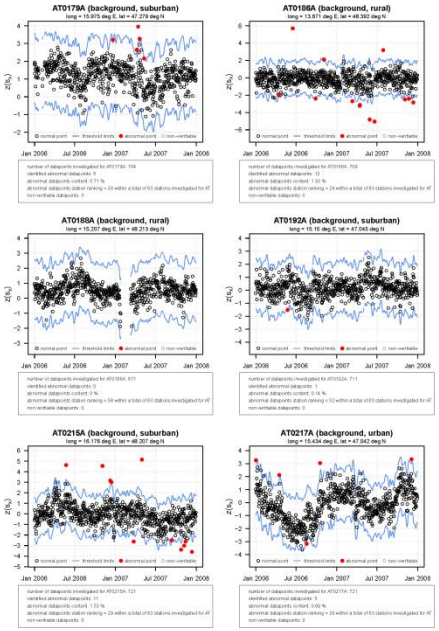
(e.g. θ_{ref} +/- a predefined threshold of 1.96)

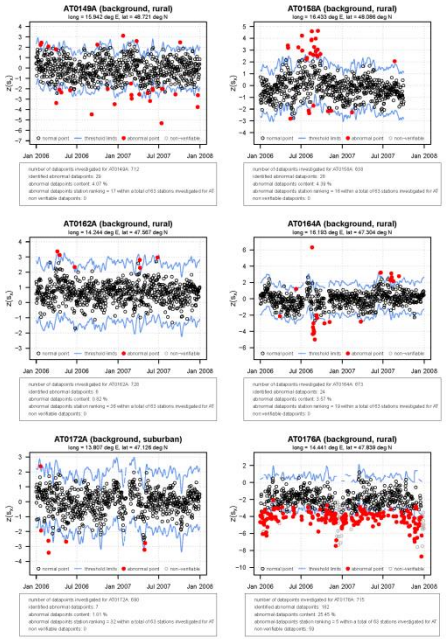
$$\left| Z_{S(x)} - \theta_{ref} \right| > \theta_{conf}$$

Automated Data Processing

- All codes prototyped in the R environment
- Directly coupled to PostgreSQL database (AirBase v.8)

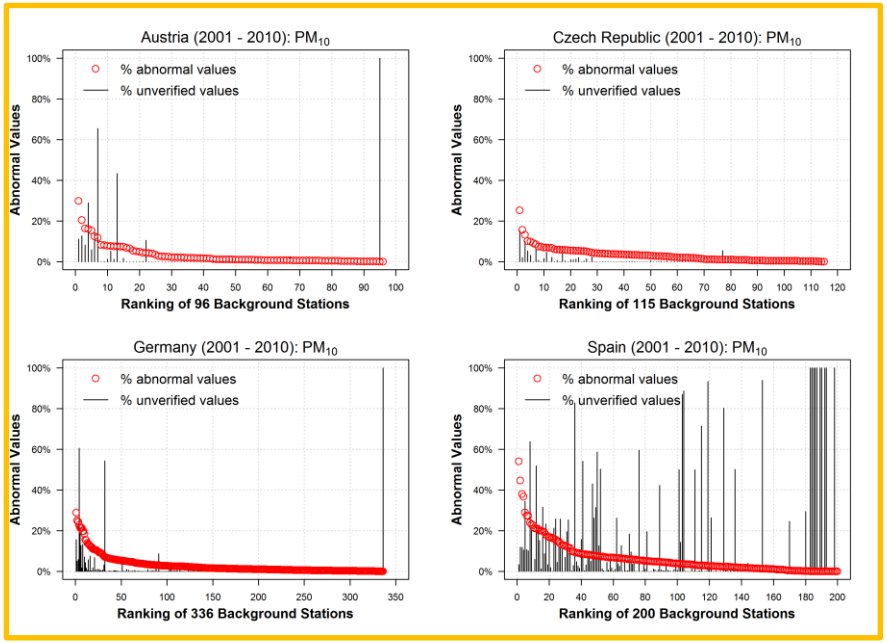


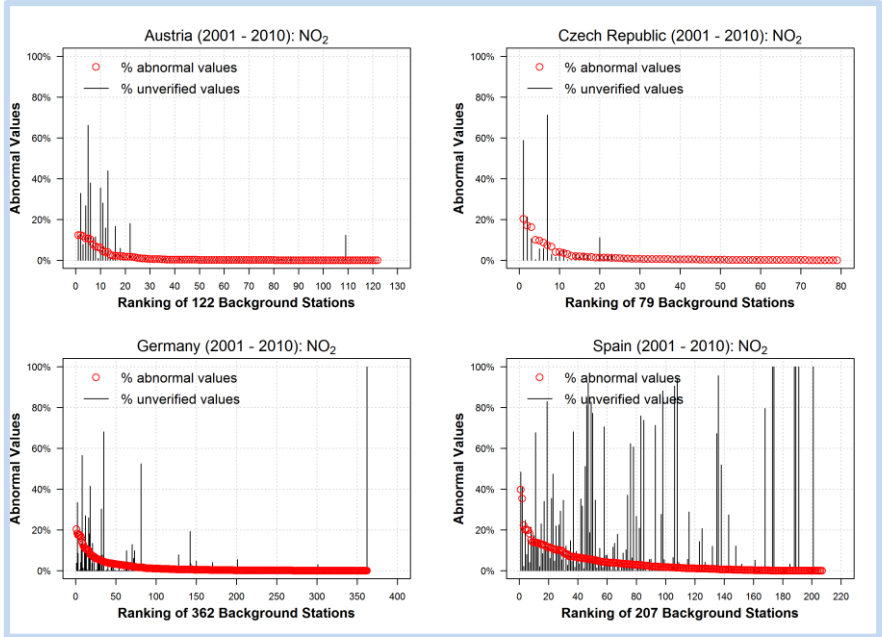




Summary of Outcomes

- 2001 - 2010 records of AirBase v.8

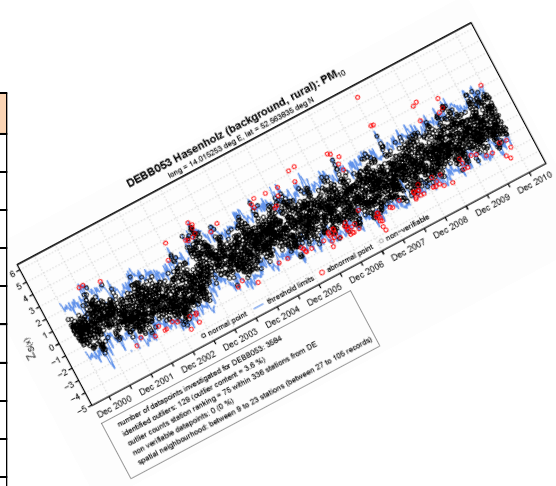




Complete 2001 - 2010 time series catalogues for 18 selected countries: PM_{10}

Screening Tool - June 2015 Version (AirBase v.8 2001 - 2010 data on PM_{10})					
Country	Stations	Records	Computation Time *	Non-Verifiable Records	Abnormal Values
AT	96	210531	31.6 min	3.2%	3.4%
CZ	115	261072	42.8 min	0.7%	2.9%
DE	336	794383	125.8 min	1.1%	2.8%
ES	200	297588	26.6 min	15.7%	6.2%
FR	341	763070	90.8 min	4.1%	6.3%
GB	70	155887	12.1 min	22.7%	4.5%
IT	250	334593	45.3 min	6.5%	4.5%
NL	33	74359	6.7 min	1.8%	5.7%
BE	44	87686	11.6 min	1.0%	3.7%
BG	33	60404	3.8 min	7.9%	8.7%
CH	22	68045	5.2 min	10.7%	7.4%
FI	12	26821	1.6 min	88.2%	1.9%
HU	15	33141	1.9 min	23.8%	8.2%
MK	2	2220	0.2 min	100.0%	0.0%
PL	277	365578	41.1 min	3.7%	5.1%
PT	46	80811	7.6 min	10.6%	5.0%
RO	43	27801	2.1 min	50.5%	6.9%
SK	29	60525	4.1 min	9.3%	5.2%

* test run with 12 cores on Intel Xeon X5680 3.33 GHz CPUs with R version 3.0.2 running on Windows-7 64-bit

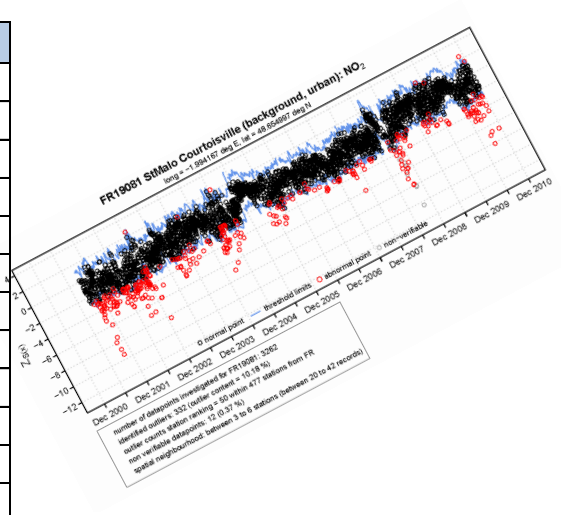


European Commission

Complete 2001 - 2010 time series catalogues for 18 selected countries: NO_2

Screening Tool - June 2015 Version				(AirBase v.8 2001 - 2010 data on NO_2)	
Country	Stations	Records	Computation Time *	Non-Verifiable Records	Abnormal Values
AT	122	325652	71.0 min	4.1%	1.4%
CZ	79	211584	28.6 min	3.3%	1.9%
DE	362	928205	166.0 min	2.1%	1.5%
ES	207	370518	34.0 min	17.2%	3.7%
FR	477	1179662	198.4 min	4.1%	3.5%
GB	98	223987	25.7 min	14.2%	4.0%
IT	324	554506	101.2 min	7.0%	2.0%
NL	45	111204	14.2 min	0.3%	2.9%
BE	40	101372	15.5 min	1.0%	2.5%
BG	18	28865	2.0 min	36.0%	5.6%
CH	26	87105	8.1 min	11.1%	3.1%
FI	15	39011	2.6 min	81.7%	1.4%
HU	15	37141	2.5 min	42.5%	4.5%
MK	2	2308	0.3 min	100.0%	0.0%
PL	310	336545	42.6 min	4.3%	4.5%
PT	48	97948	10.3 min	14.7%	2.9%
RO	60	58102	4.5 min	34.2%	9.2%
SK	24	41329	3.1 min	19.4%	7.1%

* test run with 12 cores on Intel Xeon X5680 3.33 GHz CPUs with R version 3.0.2 running on Windows-7 64-bit



Joint Research Centre

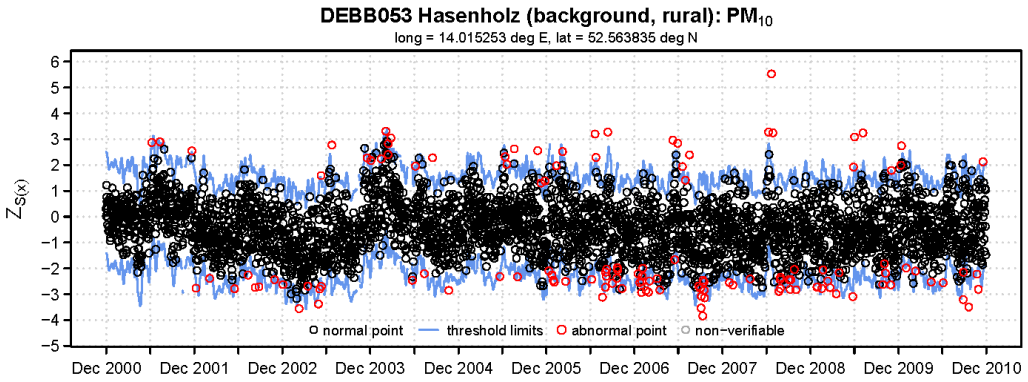
How to use this information?

Screening Tool - June 2015 Version				(AirBase v.8 2001 - 2010 data on NO_2)	
Country	Stations	Records	Computation Time *	Non-Verifiable Records	Abnormal Values
AT	122	325652	71.0 min	4.1%	1.4%
CZ	79	211584	28.6 min	3.3%	1.9%
DE	362	928205	166.0 min	2.1%	1.5%
ES	207	370518	34.0 min	17.2%	3.7%
FR	477	1179662	198.4 min	4.1%	3.5%
GB	98	223987	25.7 min	14.2%	4.0%
IT	324	554506	101.2 min	7.0%	2.0%
NL	45	111204	14.2 min	0.3%	2.9%
BE	40	101372	15.5 min	1.0%	2.5%
BG	18	28865	2.0 min	36.0%	5.6%
CH	26	87105	8.1 min	11.1%	3.1%
FI	15	39011	2.6 min	81.7%	1.4%
HU	15	37141	2.5 min	42.5%	4.5%
MK	2	2308	0.3 min	100.0%	0.0%
PL	310	336545	42.6 min	4.3%	4.5%
PT	48	97948	10.3 min	14.7%	2.9%
RO	60	58102	4.5 min	34.2%	9.2%
SK	24	41329	3.1 min	19.4%	7.1%

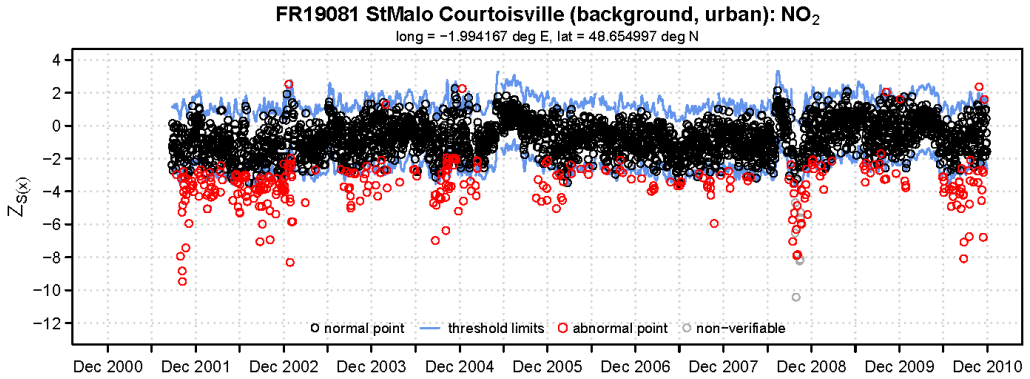
* test run with 12 cores on Intel Xeon X5680 3.33 GHz CPUs with R version 3.0.2 running on Windows-7 64-bit

Conclusions about outlier content are dependent (i) on the adjustment of the screening parameters and (ii) structural constraints stemming from the network design.

How to use this information?



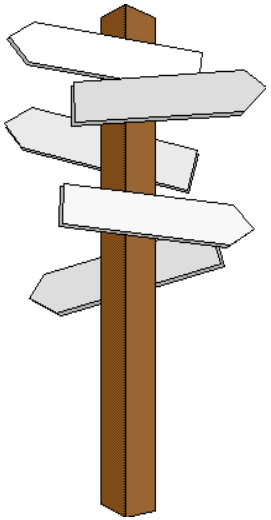
number of datapoints investigated for DEBB053: 3584
identified outliers: 129 (outlier content = 3.6 %)
outlier counts station ranking = 75 within 336 stations from DE
non verifiable datapoints: 0 (0 %)
spatial neighbourhood: between 9 to 23 stations (between 27 to 105 records)



number of datapoints investigated for FR19081: 3262
identified outliers: 332 (outlier content = 10.18 %)
outlier counts station ranking = 50 within 477 stations from FR
non verifiable datapoints: 12 (0.37 %)
spatial neighbourhood: between 3 to 6 stations (between 20 to 42 records)

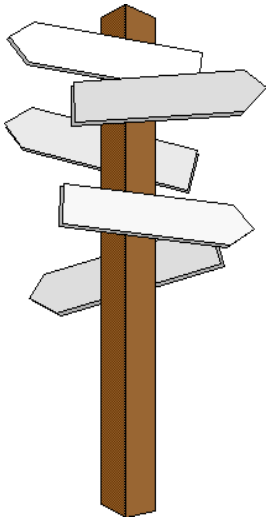
Where to use this information?

- We anticipate that the screening method can be a useful pragmatic AirBase pre- / post-processing tool for
  - AQ-Modellers (pre-screening of data selected for validation)
  - Preparation of data summaries (e.g. EEA)
  - Spatial and temporal trend analysis
  - Statistical evaluations of air quality
  - May also support QA/QC with a short feedback cycle for network operators when implemented in real or near to real time mode



Discussion

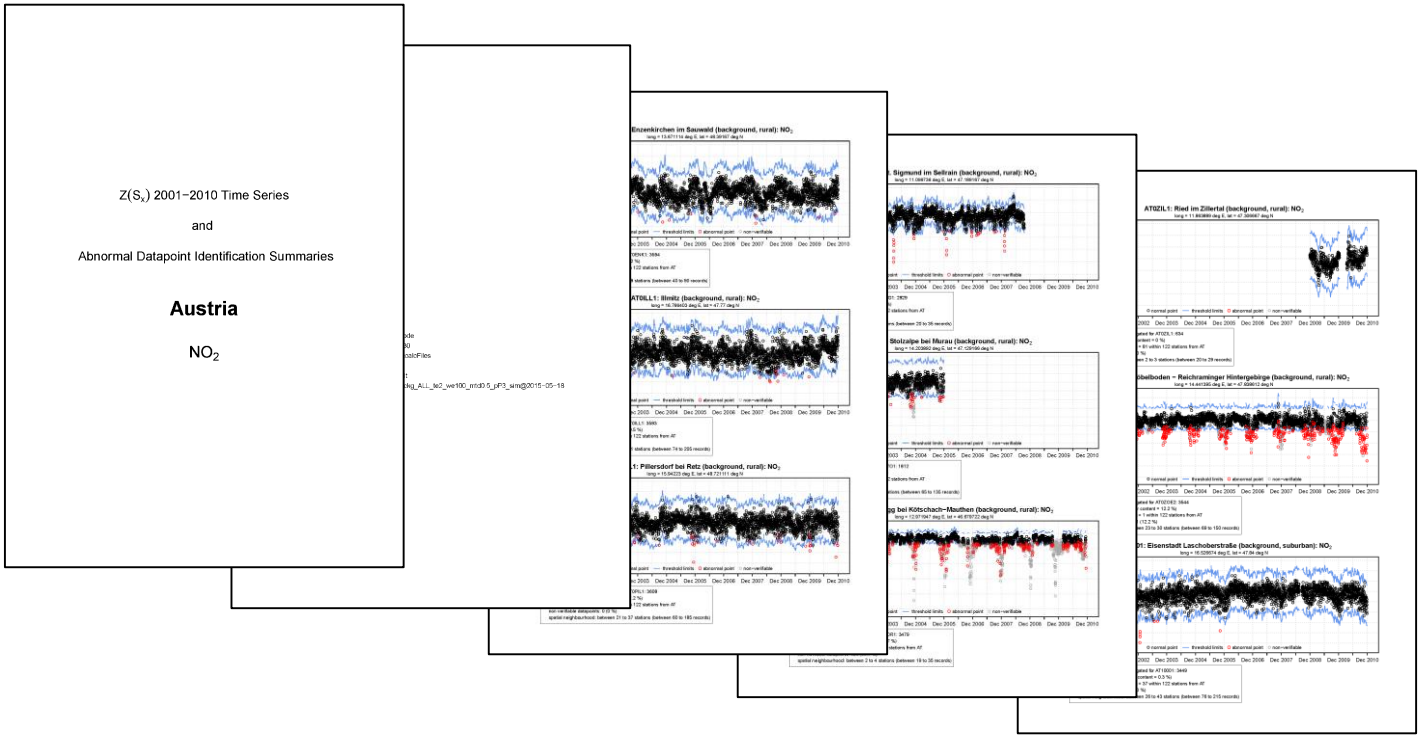
Where to use this information?



- AQ-Modellers (pre-screening of data selected for validation)
- Preparation of data summaries (e.g. EEA)
- Spatial and temporal trend analysis
- Statistical evaluations of air quality
- May also support QA/QC with a short feedback cycle for network operators when implemented in real or near to real time mode
- The primary research interest might **often** be **directed towards the anomalies themselves**. For example, an outlier detection method can be used as a tool to identify irregular emission events.

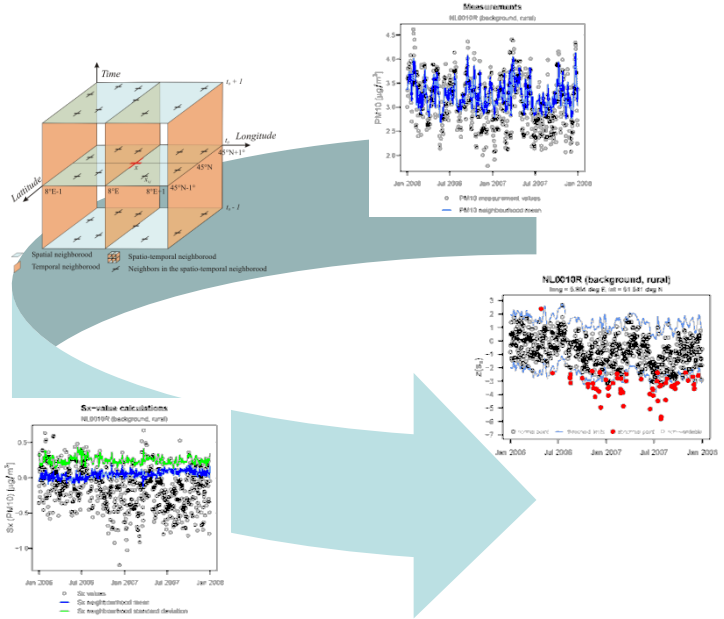
How to make this information available?

Downloadable catalogues & time series from the FAIRMODE homepage?



European Commission

Thank you for your attention!



Questions and Suggestions?