

Measurement validation by observation predictions

*Estimating observations for additional
quality control of air quality measurements*



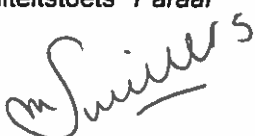




National Institute for Public Health
and the Environment
Ministry of Health, Welfare and Sport



Measurement validation by observation predictions

*Estimating observations for additional quality control of air
quality measurements*

<p>Kwaliteitstoets <i>Paraaf</i></p>  <p>Naam <i>Saskia Willers</i></p>  <p>Naam <i>Joost Wesseling</i></p>	<p>Autorisatie <i>Paraaf</i></p>  <p>Naam <i>Marcel Koeleman</i> Functie <i>Bureauhoofd Lucht</i></p>
---	---

Auteur (s): Sef van den Elshout, André Snijder (DCMR)
Lan Nguyen, Ronald Hoogerbrugge (RIVM)

Afdeling :Expertisecentrum
Bureau :Lucht
Documentnummer :21639947
Datum :15 januari 2014

DCMR Milieudienst Rijnmond
Parallelweg 1
Postbus 843
3100 AV Schiedam
T 010 - 246 80 00
F 010 - 246 82 83
E info@dcmr.nl
W www.dcmr.nl

Contents

Summary	5
1 Introduction	6
1.1 Why additional quality control	6
1.2 Study area data and scope	6
1.3 Scope of the study	7
2 Comparing expected and measured air quality observations	8
2.1 Modelling expected observations, variables used	8
2.2 Modelling approaches	8
2.3 Prediction modelling with validated and unvalidated data	10
2.4 Statistical screening	10
3 Results	11
3.1 Screening datasets Schiedamsevest (urban background)	11
3.2 Screening datasets A13 (motorway station)	16
3.3 Data quality and model specification	20
3.4 Stability of models over time	21
3.5 Testing for discontinuities after maintenance	23
3.6 General applicability of the model approaches	25
4 Discussion and conclusions	28
4.1 Introduction	28
4.2 Validation support and early warning - general concepts	29
4.3 Interpretation of results	30
4.4 Generalisation of results	32
4.5 Conclusions	33
5 Literature	34
Annexes	35
A1 PLS modelling: additional trials using unvalidated data and inverse wind speed	35
A2 Additional information model stability	37
A3 Results using distant background stations (OLS models)	40

Summary

The validation of air quality measurements is guided on the one hand by automatic logs of equipment malfunctioning and on the other by the expert judgement interpretation of occurring trends, extreme values, etc. Generally this works quite well.

The question arose if the process of 'expert judgement' of the likeliness of a certain measurement value could be facilitated and further formalised. This can be done by predicting the measured value by using a (statistical) model. If the model provides adequate predictions, the predictions can be used as a yardstick to appraise the measurements. Unlikely measurements could hint at changes in the sources of air pollution or measurement error. In both cases additional attention to the observed data is warranted.

This document describes and compares techniques to predict measurements using a number of case studies. The results show that - if sufficient reference data is available to develop prediction models – validation can be facilitated using these statistical techniques. Their use will result in easier, and potentially more uniform, validation and in better metadata. This is particularly relevant for studies of the behaviour of sources over time.

If these techniques are used to assess hourly or daily data for outliers / deviating behaviour (e.g. to filter data before their use real-time applications) the best possible models are required. If they are mainly used in support of validation (e.g. to detect deviating trends) model quality is less important. Simple ordinary linear regression models, having the benefit of being transparent in their operation, can be used as well.

1 Introduction

1.1 Why additional quality control

Monitoring ambient air quality is surrounded by many QA procedures. Dutch monitoring networks generally comply to ISO/IEC 17025. ISO/IEC 17025 indicates various technical procedures and tests to monitor the correct operation of the monitoring system such as the use of reference materials (section 5.9.1.a). In addition it also suggests the 'correlation of results' as an additional method to assess the correctness of the measurements (section 5.9.1.e.). In this document a method is presented to assess the likeliness of a measurement result as additional information to support the validation process.

Unexpected measurement results could indicate that something has changed in the sources of air pollution that influence the observations, e.g. a (temporary) closure of a road, or that something went wrong without obvious technical failures. In the first case additional meta data could be warranted to explain (for future use of the data) the deviating measurements. In the second case the data might have to be rejected after all. Increasingly, monitoring equipment is monitored by remote control and the periods of autonomous operation increase. Although this is highly desirable from an operational perspective, it also implies that visual inspection on the ground of both the monitor and the conditions in the surroundings of the monitoring station are being reduced. We developed a method to support the assessment of the 'likeness of the correctness of the observations' and tested it on a number of datasets. The results show that questionable data still occurred in validated data sets, demonstrating the usefulness of this additional statistical validation.

The approach presented in this document is based on estimating each measured value and comparing the measurement with its estimate. The concept of comparing measurements with modelled estimates was described in Carslaw and Carslaw (2007). The difference between the two, the residuals of the estimation model, are analysed for deviances (patterns, outliers) indicating unexpected measurement results. Carslaw and Carslaw used the approach to study changes in vehicle direct NO₂ emissions. Whilst doing this study they discovered an anomaly in an observed time series that coincided with a maintenance event and suggested that the described approach could be used for quality control. In this study we built on this concept though the models to make the predictions are different.

1.2 Study area data and scope

The present study is conducted in Rijnmond (the Rotterdam port-industrial area). DCMR, the regional EPA runs an air quality monitoring network in this area. The Dutch national monitoring network (run by RIVM, National Institute for Public Health and the Environment) also has a number of sites in Rijnmond. This report focuses on NO₂ though the concepts used can be applied to other pollutants as well. RIVM and DCMR closely collaborate in the field of air quality monitoring and amongst others jointly operate a monitoring station where most pollutants are monitored in duplicate to assure comparability of the monitoring results of both networks.

Developing and testing of the methods described was done on monitoring data for the period 2011 and 2012 with emphasis on a roadside monitoring site and an urban background site. An additional scan of all NO₂ monitoring stations and of historical data will be performed by DCMR. Some examples can be found in the annexes.



Figure 1. Study area and monitoring locations. The two main case study locations are Schiedamsevest (urban background) A13 Overschie (motorway). Other stations shown (and some not shown) were used to develop the statistical models. Meteo was derived from Rotterdam airport, approximately 1 km north of the A13 monitoring station.

1.3 Scope of the study

The methods developed to predict hour-by-hour air quality measurements can be used for a number of purposes.

- Testing the likeliness of measured observations as support for data validation (the main reason to develop this method)
- Early detection of potential problems at monitoring stations to avoid, in case of a problem, having to reject large amounts of data during validation.
- Missing data can either be due to technical problems or to a non continuous monitoring strategy. NO₂ has distinct diurnal and seasonal patterns. Missing data affect the uncertainty of observed mean, and non-random missing data will create bias. RIVM uses a method to estimate missing regional background daily average PM data (Mooibroek, 2013). DCMR employed a method to correct potential bias in incomplete NO₂ data series before (Elshout, 2003). The models developed here could supplement/replace the existing methods.
- Monitoring equipment is guaranteed to work properly within a certain range of conditions. Manufacturers increasingly add sensors to their equipment providing diagnostic information. The question arises which diagnostic signals indeed lead to wrong observations (observation rightly rejected during validation) and which are less important. After all rejecting an observation also leads to increased uncertainty in the observed mean value.

The focus of this document will be on the first two bullets.

2 Comparing expected and measured air quality observations

2.1 Modelling expected observations, variables used

For each monitoring location (x,y) at each time (t) a measurement $M(x,y,t)$ is obtained. To assess the likeliness of $M(x,y,t)$ it is compared to an expectation $E(x,y,t)$. For each location and time the difference $(\Delta(x,y,t))$ between the measurement and the expectation is derived from:

$$M(x,y,t) = E(x,y,t) + \Delta(x,y,t) \quad [1]$$

Ideally $\Delta(x,y,t)$ has the following properties:

- The mean of $\Delta(x,y)$ is constant: there is no drift neither in the expectations nor in the measurements
- The mean of $\Delta(x,y)$ is 0: there is no bias between the expectations and the measurements
- $\Delta(x,y,t)$ is normally distributed, the standard deviation is small and the observations are independent (there is no pattern in $\Delta(x,y,t)$)

There are two ways to make the expectation $E(x,y,t)$:

- By looking forward where for each monitoring site (x,y) , $E(t+1)$ is a function of the previous measurements $M(t)$. This was done for example by Carslaw and Carslaw (2007) using a GAM model.
- By looking 'sideways' in space where for every time (t) , $E(x,y)$ is a function of the measurements at other monitoring sites $M(x_i,y_i)$. This was used by RIVM (Mooibroek, 2013) to estimate missing $PM_{2.5}$ values.

We opt for the spatial approach where $E(x,y,t)$ depends on $M(x_i,y_i,t)$ at other monitoring sites in the same region as this is more likely to be able to capture gradual drift. In addition to the other measurements meteorological variables are used. $E(x,y,t)$ is modelled using multiple linear regression (see section 2.2.1) and a principal component technique (see section 2.2.2).

2.2 Modelling approaches

2.2.1 Linear regression approach

Linear regression models (ordinary least squares – OLS) were built using an average background concentration - $B(t)$, meteo and a number of other variables. Matlab interactive stepwise regression was used (but this could even be done in a spreadsheet) and variables significant at $p = 0.05$ were retained in the model. Hourly measurement and meteorological data are used. The meteo data is from Rotterdam airport.

The variables used in the OLS approach are:

- $B(t)$ that is calculated as the average measured concentration at time (t) of 5 background stations in the area. When the prediction method is applied to one of the background stations that specific station is excluded and $B(t)$ will be based on the other 4 stations.
- A dummy variable that differentiates working and weekend days $W(t)$.
- Six dummy variables indicating wind direction in groups of 60 degrees - $WR1(t)$ to $WR6(t)$. $WR1$ represents a wind direction between 0-60 degrees each consecutive WR variable covers an additional 60 degrees. The situation where the wind direction could not be established is the default situation (to avoid over-fitting).
- The inverse wind speed - $WS(t)$ is used as a measure for exposure conditions.
- Precipitation – $P(t)$ and Temperature – $T(t)$

$$E(x,y,t) = \beta_0 + \beta_1*B(t) + \beta_2*W(t) + \beta_3*WR1(t) + \beta_4*WR2(t) + \beta_5*WR3(t) + \beta_6*WR4(t) + \beta_7*WR5(t) + \beta_8*WR6(t) + \beta_9*T(t) + \beta_{10}*P(t) + \beta_{11}*(WS(t)+1)^{-1} \quad [2]$$

For each monitoring site a regression model is developed and a time series of expected values is calculated that can be compared with the actual measured values. The hourly differences between measured and expected values $\Delta(x,y,t)$, are calculated and plotted and analysed. See chapter 3.

The models are evaluated by their R^2 -adjusted and RMSE parameters. Note that RMSE is not the regular RMSE output from the regression tool. To be able to compare regression RMSE with PLS RMSE both are calculated as:

$$RMSE = \sqrt{(\sum(\text{model} - \text{measurement})^2/n)} \quad [3]$$

2.2.2 PLS regression approach

PLS (partial least squares regression) is a modelling technique similar to PCA (principal components analysis¹). Instead of regressing a property (in this case, a concentration) onto a set of variables, the property is regressed onto the score of principal components of these variables. In this study a dataset of 23 variables was used (for a two year data set this is an array of 17544 rows and 23 columns). These variables are:

- NO₂ hourly concentrations measured at 18 individual monitoring sites in Rijnmond and surrounding area. All monitoring sites in this area were used regardless of their classification.
- 5 meteo variables: wind speed, temperature, precipitation and 2 variables indicating wind direction. The wind direction is not used as such but the vectors of the wind direction were used. The meteo variables were multiplied with a factor to get values of the same order as the concentration: $WS(t)*10$, $T(t)*10$, $P(t)*20$, $\sin(WR(t))*20$, $\cos(WR(t))*20$.

The PLS modelling was performed with a PLS_Tool box² that can be run within the MATLAB environment. From the original parameters this tool calculates principal components, performs regression with the chosen principal components (5 in this study) and gives predicted values. Some trials show that auto-scaling of data (a function of the PLS_Tool)³ prior to regressing improves the results of the models. Therefore this pre-processing was applied to all runs. The results reported here were obtained with 5 principal components. The used dataset has a few percent missing data. Because the PLS tool can not deal with missing data, prior to modelling these data were filled with values estimated from other data (see also appendix A1).

PLS regression is particularly suited when the matrix of predictors has more variables than observations (not relevant in this case) and when there is multicollinearity among the variables in the model. Standard (OLS) regression will fail in these cases. The OLS models in this document overcome the multicollinearity problem of using several background stations by averaging them into a single background variable. Inevitably some information is lost in this process. The PLS regression is more efficient in using the information of all the monitoring stations in the sample.

When the text refers to 'OLS' or 'PLS', the modelling approaches as outlined in this and the previous section are meant. **It should be noted that apart from the regression techniques, both approaches also differ (slightly) in the variables used and the way the variables are handled** (e.g. individual monitoring stations <> average of background).

¹ PCA was successfully used in Nguyen et al. (2012) to characterise the behaviour of Dutch air quality monitoring stations

² See also <http://www.eigenvector.com/>

³ i.e. mean-centring and scaling the column to unit variance

2.3 Prediction modelling with validated and unvalidated data

The prediction methods as described here were initially developed on validated data. If this statistical screening is to be turned into an operational tool assisting in the validation process, it has to work on unvalidated data as well. In this study we developed the models using a two year dataset (≈ 17000 hourly values) so with a typical validation 'loss' of less than 5% it is not expected that this will affect the models significantly. Some trials with the PLS approach show that models developed on validated and unvalidated data give the same results (see appendix A1).

If the unvalidated data contains substantial errors, this might affect the model. As it turned out the test data did contain periods, that in retrospect had an extended period of data that should not have passed validation. This provided an opportunity to see how sensitive the models are to measurement error in the dataset. The tests developed for deviating data were also used (in one case) to see if the statistical screening presented here reaches the same conclusion as the manual validation process (see section 3.1.4).

2.4 Statistical screening

The differences between the measured and the predicted value are interpreted on different averaging times. Test criteria were developed for hourly and daily (24h moving average) results of the differences. This could be helpful in the early detection of deviating measurements. Additional tests were developed for longer averaging times (1, 2 and 4 weeks) to see if deviating trends can be detected. In all cases moving averages of the desired period were calculated and RMSE for each averaging time was determined. Testing was with 2, 3, and 4*RMSE as thresholds.

The outcome of the screening is an indication of the **likeliness of an observation**. If an observation is flagged it means that it requires additional attention during validation; it doesn't necessarily mean that something is wrong. There can be external circumstances (change of sources, change of traffic near a roadside station, etc.) that cause the observations to deviate from the expectations. The fact that measurements are flagged should lead to closer inspection of the measurements and the circumstances under which they occurred. This can reveal reasons to discard the data as wrong, or on the contrary confirm that they are deviant but correct. If indeed the sources that influence the station have changed, and the change is permanent the prediction model would have to be run again to create new expectations $E(x,y,t)$.

3 Results

The methodology was development on data from the urban background station Schiedamsevest in Rotterdam centre and the A13 motorway station in Overschie (Rotterdam north). In view of the currently developed screening method, both monitoring sites happen to have periods where anomalies seem to occur and/or periods that should not have passed validation.

3.1 Screening datasets Schiedamsevest (urban background)

3.1.1 Model characteristics

The regression model and its characteristics are shown in table 1. The performance characteristics of the PLS results are shown as well. The PLS approach seems to perform better than the OLS approach though the differences are not very large.

Table 1: Schiedamsevest – modelling approach performance (OLS and PLS) and OLS coefficients

	OLS coefficients	PLS model
Background concentration	0.91	
Weekday (W)	2.26	
Wind direction WR1	-8.60	
Wind direction WR2	-2.12	
Wind direction WR3	1.84	
Wind direction WR4	4.72	
Wind direction WR5		
Wind direction WR6	-4.02	
Temperature	-0.03	
Precipitation	-0.13	
(Windspeed+1) ⁻¹	4.65	
Intercept	3.70	
RMSE	9.564	7.451
R2_adjusted	0.763	0.845
Variables used	11	23 (5) ⁴
# Observations	17395	17395

3.1.2 Interpretation of results – visual inspection

Figure 2. shows plots of a 24-hour moving average of the differences between the model and the observations. The figures for both models are quite similar though the regression model occasionally leads to large spikes. Apart from the occasional spikes in the 24 hour moving average, both graphs show a distinct gradual upward drift from mid August 2011 until mid January 2012. From January 2012 onward the residuals show only a small variation around -2. The fact that the residuals don't oscillate around 0 is a consequence of the upward drift earlier in the sample period that was used to fit the models. Since the modelling leads to residuals with an average of 0, substantial deviations in the sample period that is used to build the model leads to a small bias in periods when the data are correct. For visual inspection this is not a problem but if automatic flagging of results is introduced this needs attention.

⁴ Resulting in 5 principal components that make up the actual prediction model

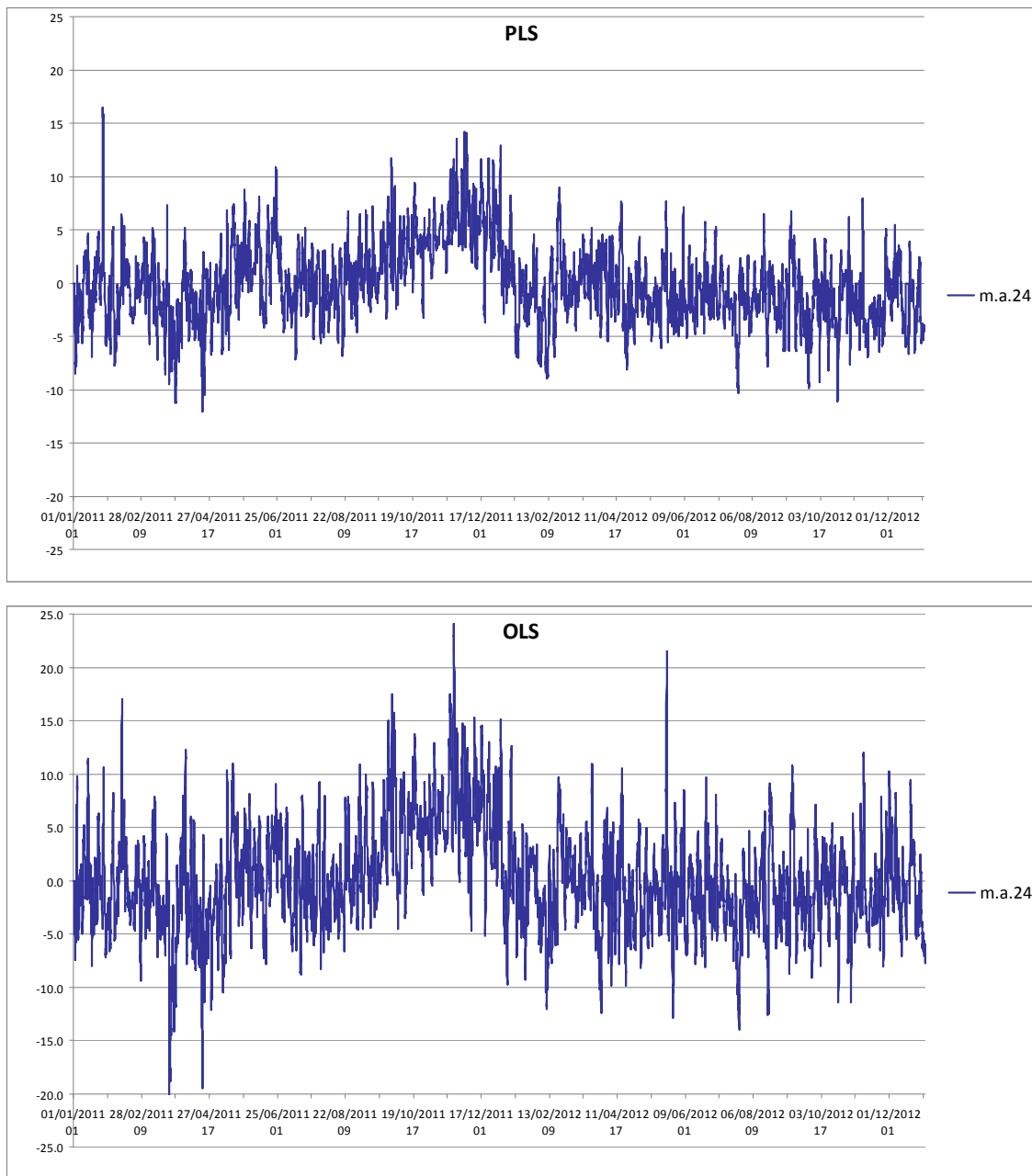


Figure 2: Schiedamsevest – 1 day moving average plots of the differences (measurement-model)

The PLS graph for the 24h moving average displays less variation and misses a few extreme spikes that are found in the OLS graph. Apparently the PLS model is better capable of capturing the hourly and daily variation at the monitoring station. The R²-adjusted and RMSE scores of PLS approach are better than those for OLS approach but the graph shows that this is (partially) the result of less spikes and not just overall smaller random noise. This is an advantage of the PLS as the spikes are used to flag potentially erroneous measurements.

Figure 3. shows the weekly averaging times for both models including the 1 and 2*RMSE bands. Both models agree fairly well. If a 2*RMSE threshold is used to flag potential deviations the visual inspection shows that 4 and 6 periods are flagged by the PLS and OLS models respectively. If 1*RMSE is used for flagging the results are 19 and 18 respectively. The 1*RMSE criterion seems to produce a substantial number of suspect periods though this is somewhat inflated by the period with the upward drift causing all results to be more extreme.

The 2*RMSE criterion flags an operationally acceptable number of periods though it is quite late in positively identifying the gradual upward trend. In this particular case one would lose an additional 5 to 6 weeks in diagnosing a potential problem.

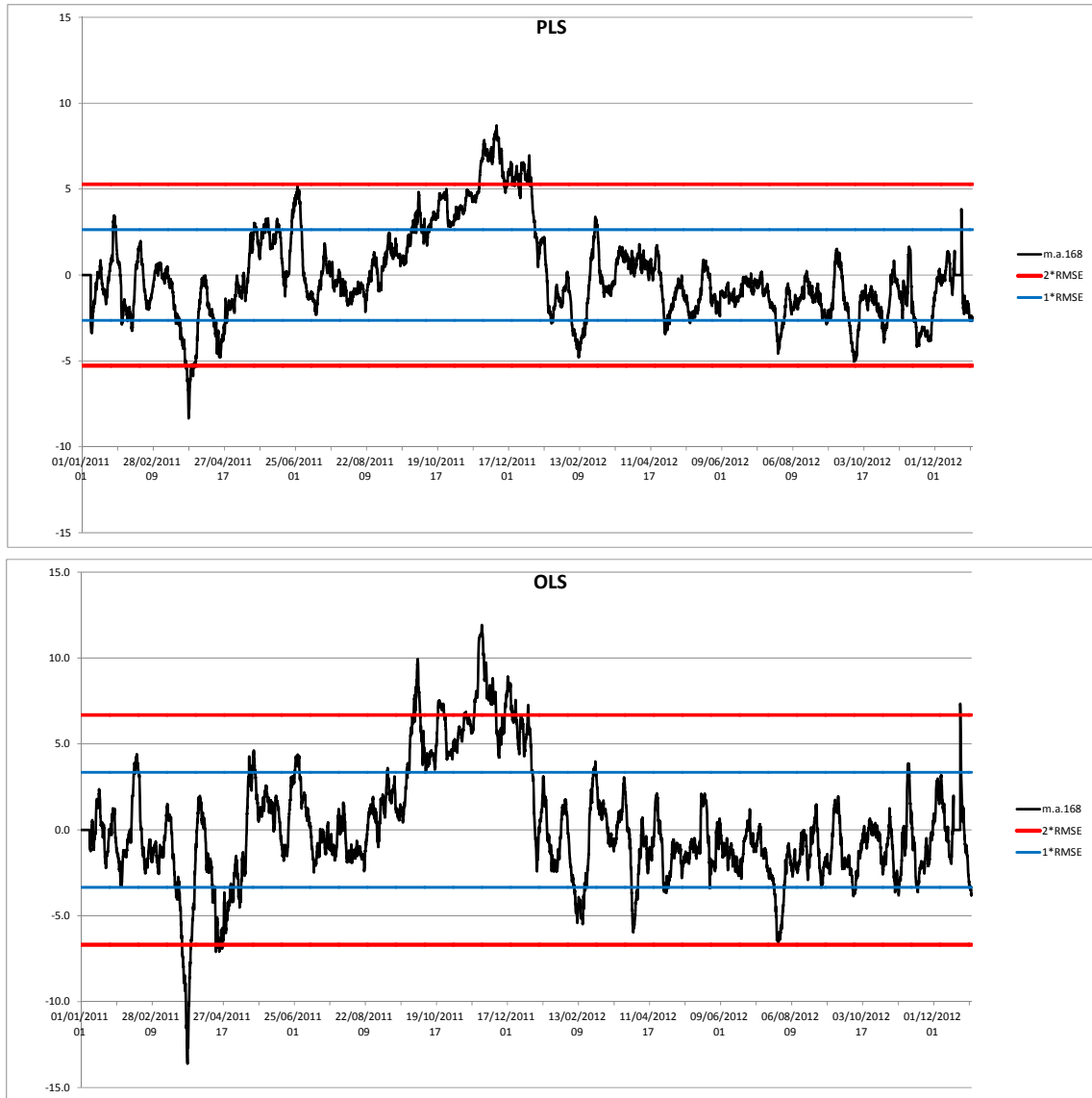


Figure 3: Schiedamsevest – 1 week moving average plots of the residuals

Figure 4. shows the 4-week moving average with 1*RMSE bands. The 4-week moving average flags only a few periods and, as was remarked before, the hits in the second year are mainly due to the bias that was created in the model by the deviation in the first year. The 4-week moving average would flag the upward drift only a few days after the weekly moving average with a 1*RMSE criterion. The 4-week moving average therefore has the advantage of offering a low threshold and relative early detection without the potential disadvantage of numerous periods being flagged.

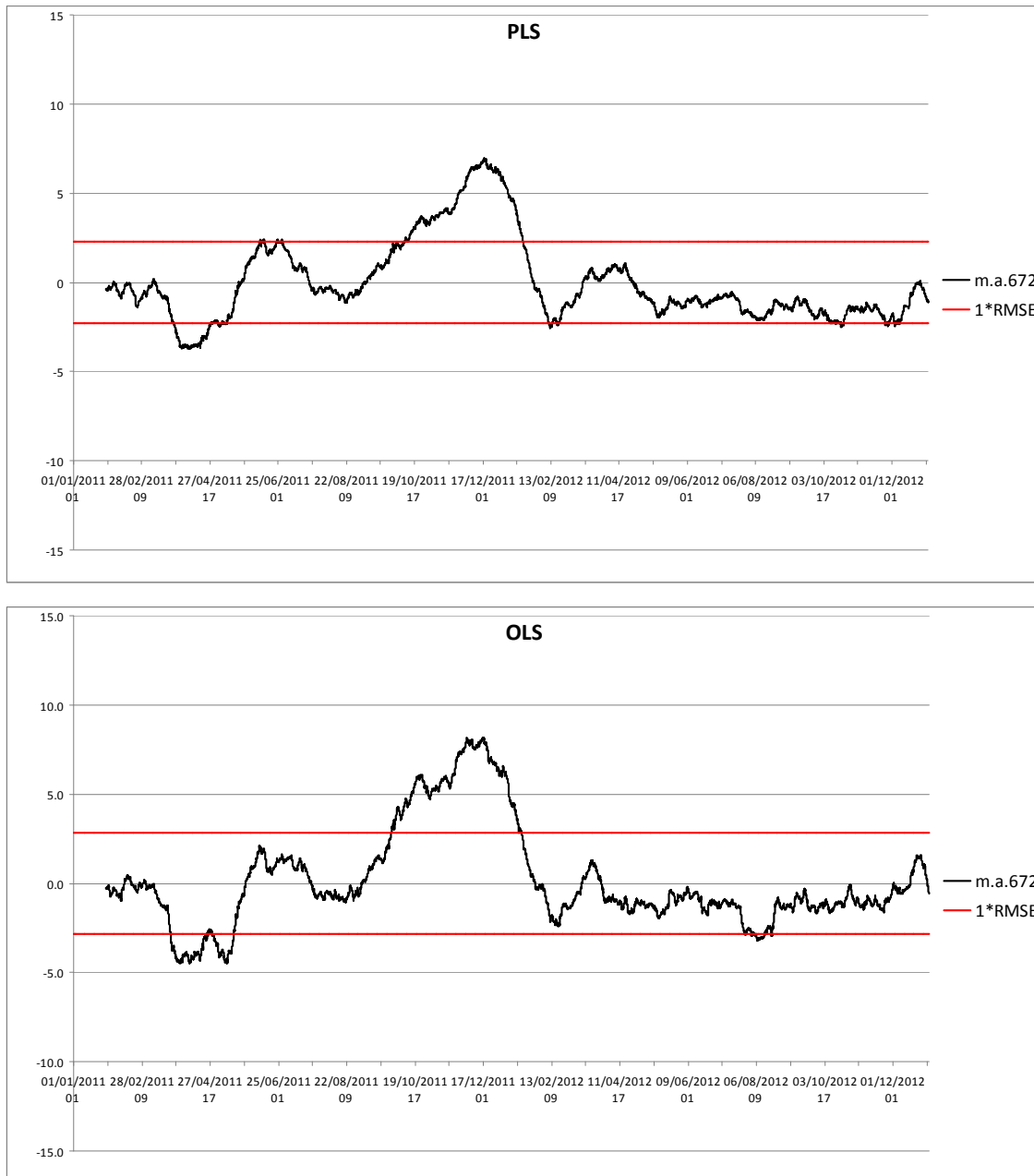


Figure 4: Schiedamsevest – 4-week moving average plots of the residuals

3.1.3 Criteria for automatic screening

For an automatic screening, test criteria to flag potentially deviating measurements are needed. RMSE was calculated for the various averaging times and thresholds for 1, 2, 3 and 4*RMSE were tested. The results for both models are shown in table 2. The table shows that RMSE rapidly drops with increasing averaging time. From a week onward, this decrease is small. For short averaging times 1 and 2*RMSE thresholds result in a substantial number of suspect hours needing additional attention. This is not desirable. For hourly data even the 3*RMSE threshold would lead to many hours being flagged. However, this is partly due to the period with upward drift. The 3*RMSE criterion is probably not practical for hourly values and 4*RMSE should be used. For the 24 hour moving average 3*RMSE seems suitable.

Table 2: Number of hours or periods (in case of the moving averages) when the specified RMSE criteria is exceeded by the models

#*RMSE	PLS				OLS			
	1 hour	24 hour	1 week	4 weeks	1 hour	24 hour	1 week	4 weeks
1	3955	327	56	30	4052	314	44	13
2	838	68	20	3	928	68	18	3
3	247	13	8	1	268	13	2	0
4	88	1	0	0	86	3	1	0
RMSE	7.45	3.67	2.64	2.31	9.56	4.92	3.35	2.82

The 4-week moving average is best used for visual inspection as shown in the previous section. Flagging based on 1*RMSE leads to quite a number of hits with the PLS model according to table 2. However, this is somewhat misleading if the table and figure 4 are compared. There happens to be a lot of wavering of the moving average around the 1*RMSE threshold causing a substantial number of hits. The number of 'hits' with the OLS model is much lower and happens to do more justice to the situation in this case. Increasing the threshold to 2*RMSE based on these results is not recommended, it will delay the detection of the drift period substantially. A combination of 2*RMSE for weekly moving averages + 1*RMSE for 4-weekly moving averages would lead to an acceptable level of warnings without losing the capacity to early detect drift issues.

There is a fair amount of agreement as far as the overall scores are concerned. A second test was done to verify if the models both pick the same hours and days. Both models agree on identifying the measurements at the end of the period of upward drift as questionable as can be seen from figure 4. The detailed results on an hour by hour basis are shown in table 3. For the short averaging times, there is some agreement but the results are not impressive. For the weekly criterion the results are slightly better. Apparently the timing of the models is slightly different. Though the overall agreement between the models is good as can be seen from the graphs, agreement⁵ on the discrete decision-making (yes or no > threshold) at a given hour only occurs in the more extreme circumstances. It also shows that the longer the averaging time, the better the agreement between the models.

Table 3: Number of hours flagged by one or both models for various combinations of RMSE threshold and averaging time.

	Hour 4*RMSE	Day 4*RMSE	Day 3*RMSE	Week 2*RMSE
Both models flag	42	0	16	713
Only PLS flags	46	13	99	410
Only OLS flags	44	23	120	370
No flag	17412	17508	17309	16051
% agreement	32%	0%	7%	48%

3.1.4 Manual and 'automatic' validation compared

All results reported so far are based on unvalidated data. This provides the opportunity to compare the flags by the models to the decisions made during the normal manual validation process. Table 4 shows the results. Looking at the 3 and 4*RMSE there is agreement between the models and the manual validation in about 13 to 25 % of the hours. During manual validation the period with the very gradual upward drift was noticed but the measurements were

⁵ Agreement is defined as the number of hours that both models flag an hour/the total number of hours flagged. Of course, if no hours are flagged (the vast majority of the time) the models also agree but this is not a very informative parameter.

accepted as drift was seen on a previous occasion. Had this period been rejected⁶ this would have improved the agreement somewhat (a few %).

Table 4: Hourly observations judged by ‘manual validation ‘ compared to the exceedence of modelled thresholds (PLS/OLS)

Threshold (RMSE)	2	3	4	Remark
Total numbers of hours flagged	838 / 928	247 / 268	88 / 86	
Hours flagged and manually rejected	51 / 49	48 / 34	28 / 21	Agreement between model and manual interpretation
Hours flagged but not rejected	787 / 879	199 / 234	60 / 65	This includes a part of the hours in the period with upward drift
Rejected but not flagged	97 / 99	100 / 114	120 / 127	Criteria to manually reject observations are too strict???
Total number of hours manually rejected	148	148	148	Rejections only occurred on technical grounds (e.g. temp. out of range, etc.)

About two thirds of the hours that were manually rejected were not flagged by the models. Rejections occur on technical grounds, e.g. instrument parameters are out of range and the manufacturer doesn't guarantee proper functioning of the monitor. The results from this analysis seem to suggest that for some of these parameters less strict criteria are feasible. After all the monitoring results do fit in the range of expected values. In particular those hours where measurement and model match closely could be inspected to see which technical grounds caused the rejection⁷

3.2 Screening datasets A13 (motorway station)

Model development was done on validated data. This implies that the analysis done in section 3.1.4 will not be repeated for this monitoring station. Like the urban background station in 3.1. this dataset contains a period with substantially deviating data.

3.2.1 Model characteristics

The regression model and its characteristics are shown in table 5. The performance characteristics of the PLS results are shown as well. Compared to the urban background station the model performance is somewhat less. On the other hand the results are also influenced by the fact that deviating data was used to fit the models. The models were refitted excluding the erroneous data. This improved the model performance characteristics notably.

Also in this case the PLS model performs (substantially) better than the regression model. This becomes even more evident looking at figure 5 with the daily averaged differences. The OLS model displays much more noise than the PLS model.

⁶ Validation decisions are often expert judgements. Statistical methods like this can never replace these judgements but the tools provide a way to quantify the likeliness of deviations and pinpoint the timing of changes in trends. In this case the trend change coincided with maintenance.

⁷ Currently the nature of the technical rejection is not recorded so this aspect was not investigated.

Table 5: A13 motorway - modelling approach performance (OLS and PLS) and OLS coefficients

	OLS coefficients	PLS approach	OLS coefficients - corrected*	PLS approach - corrected*
Background concentration	0.85		0.99	
Weekday (W)	5.44		5.43	
Wind direction WR1	-16.76		-16.24	
Wind direction WR2	-12.54		-13.07	
Wind direction WR3	7.01		6.44	
Wind direction WR4	9.94		10.73	
Wind direction WR5	7.32		8.13	
Wind direction WR6	-4.96		-3.64	
Temperature	-0.10			
Precipitation (Windspeed+1) ⁻¹			4.35	
Intercept	10.38		4.55	
RMSE	13.014	9.415	11.93	7.676
R2_adjusted	0.64	0.81	0.71	0.88
Variables used	10	23	10	23
# Observations	17014	17018	14709	14709

* New OLS model excluding the period with erroneous data from the dataset for model development⁸

3.2.2 Interpretation of results – visual inspection

Visual inspection of the moving average of the weekly differences (figure 5, right hand side) clearly reveals a period with a significant drop of 5 to 10 $\mu\text{g}/\text{m}^3$. The daily and weekly graphs show that the drop occurs almost instantly and is different from the gradual upward drift seen at the background station in the previous case study. Return to normal is also almost instantly.. The 1*RMSE threshold (weekly moving average) is exceeded on numerous occasions particularly in the OLS model. In this case using the 2*RMSE doesn't delay the detection of the problem.

⁸ The results of the analysis and the fact that the changes (deviation and return to normal) coincided with maintenance at the monitoring station led to revalidation and rejection of the flagged data.

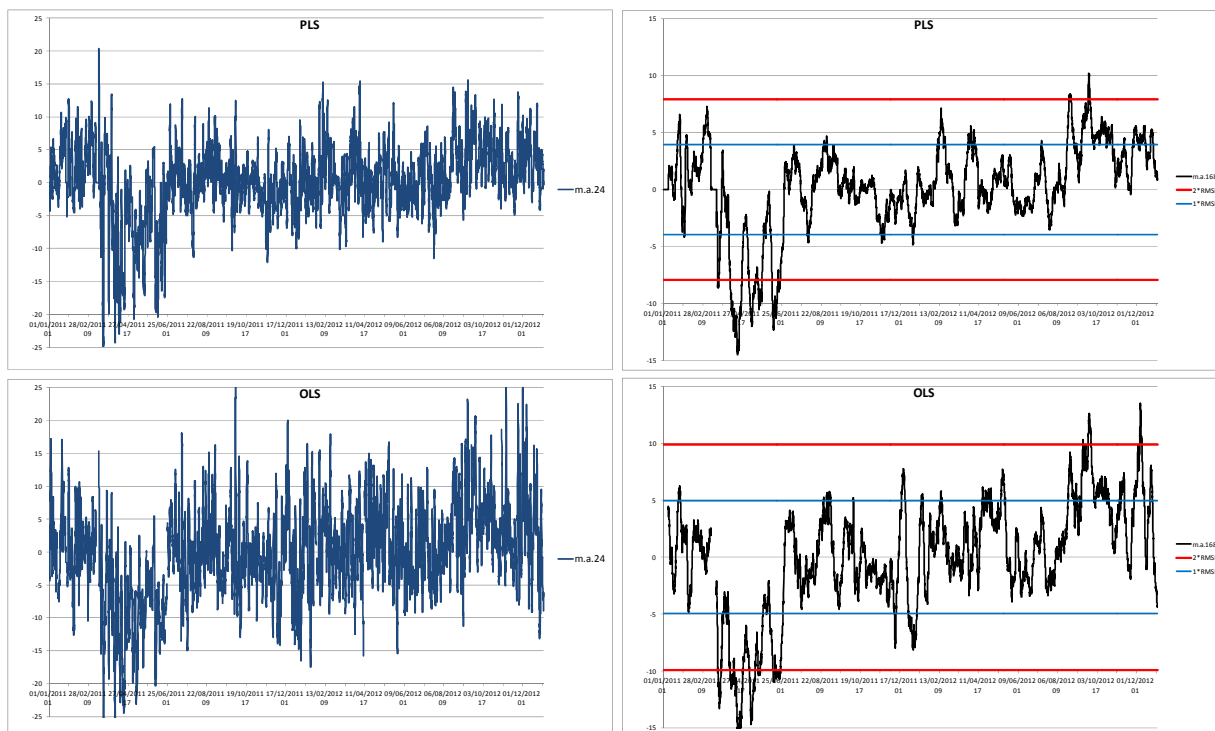


Figure 5: A13 motorway – 1 day(left) and 1 week (right) moving average plots of the residuals

Figure 6 showing the 4 week moving average shows the same broad pattern for both models. Apart from the problem in spring 2011 the models also flag a period at the end 2012. This coincides with a maintenance event.

3.2.3 Criteria for automatic screening

Results of the automatic screening for the motorway case are shown in table 6. Combinations of flagging thresholds and averaging times resulting in unrealistic small or large numbers were not considered (see table 2 for example). Given the fact that there was an extended period with erroneous measurements the hourly and daily results show a surprising small number of hits. The shorter averaging times don't seem suitable to detect the structural problems and mainly flag (potentially wrong) extreme values. It is the longer averaging times that are needed to detect structural problems.

Table 6: Number of hours or periods (in case of the moving averages) when the specified RMSE criteria is exceeded by the models (only probable combinations considered)

#*RMSE	PLS				OLS			
	1 hour	24 hour	1 week	4 weeks	1 hour	24 hour	1 week	4 weeks
1			51	17			44	14
2		56	21	5		72	19	1
3	268	24	5	0	159	14	3	0
4	85	3			30	3		
RMSE	9.415	5.525	3.972	3.329	13.014	7.252	4.963	4.150

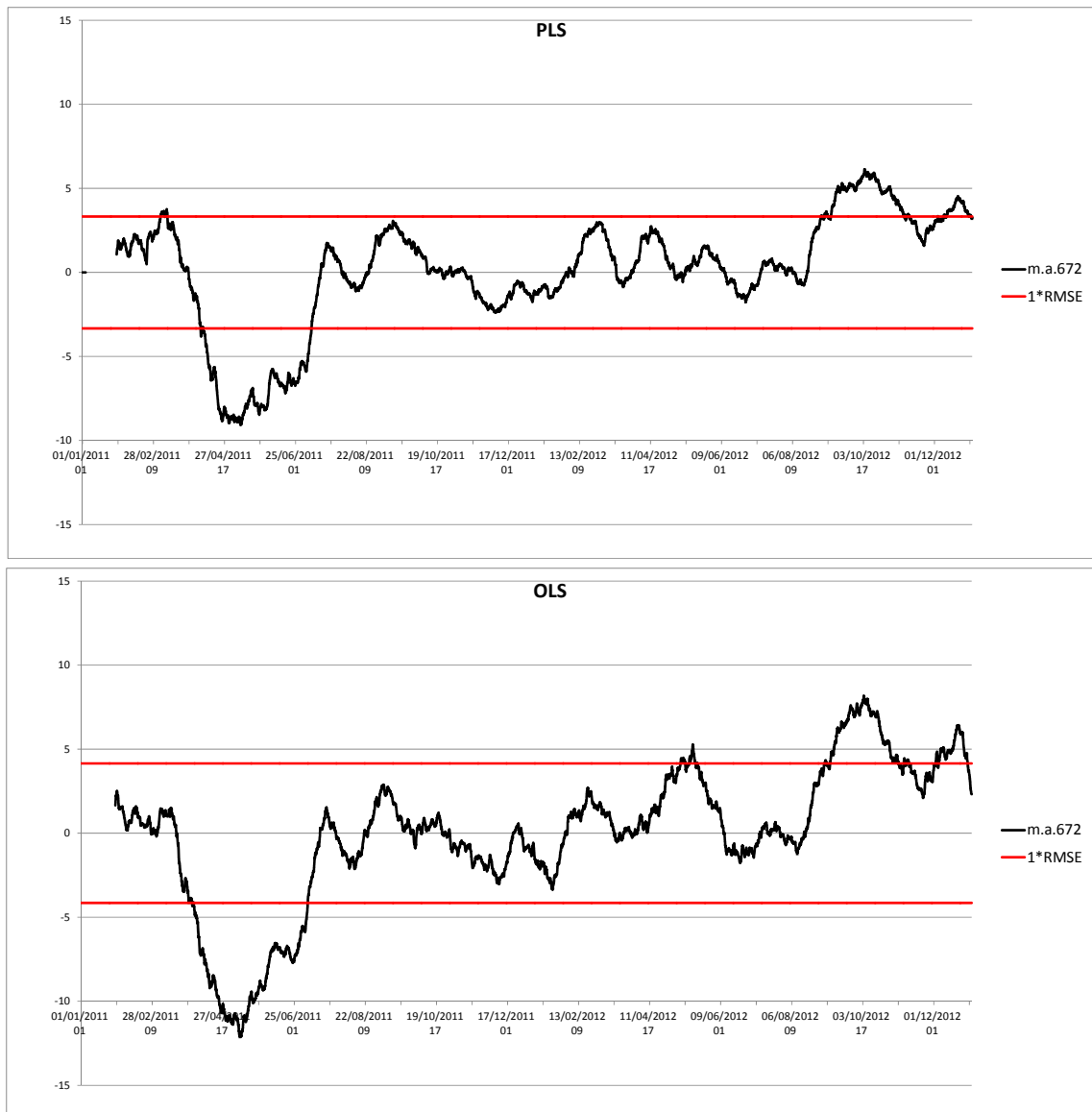


Figure 6: A13 motorway – 4-week moving average plots of the residuals

The agreement between the two models on the number of flagged hours is shown in table 7. The results are similar to the previous case. Though there is agreement on the overall trend of the differences, agreeing on the exact hour when a threshold is exceeded appears difficult (particularly when the criteria are strict and only a few hours are flagged).

Table 7: Number of hours flagged by one or both models for various combinations of RMSE threshold and averaging time.

	Hour 4*RMSE	Day 4*RMSE	Day 3*RMSE	Week 2*RMSE
Both models flag	9	3	62	758
Only PLS flags	21	3	87	477
Only OLS flags	76	31	207	478
No flag	16912	17026	16707	15486
% agreement	8%	8%	17%	44%

It should be born in mind that this dataset was validated (unlike the previous one). So table 7 could be read as: both models agree to flag an additional 3 to 62 hours as (highly) unlikely (using the shorter averaging time criteria).

3.3 Data quality and model specification

In the previous sections we saw that trouble data in the dataset used to fit the models does not affect their general operation. However, extend periods of deviating data causes two effects:

- it increases RMSE and hence reduces the sensitivity to detect problems;
- it inflates smaller deviations from 0 with the opposite sign of the problem period as the models are fit to assure that the mean of all differences equals 0; the basic concept of regression techniques.

These effects can be demonstrated by looking at the corrected models for the motorway station. See figure 7. The top graph is based on the revised OLS model (see table 5.) and the bottom the original graph as shown in figure 6.

The graph shows that the 1*RMSE band shrinks from 4.2 to 2.6 $\mu\text{g}/\text{m}^3$. The upward peaks in the second year shrink somewhat and the whole graph shifts a few tenths of a microgram downward to compensate these peaks. And whereas 1*RMSE previously seemed a practical criterion, it is rather strict when the model is developed on correct data so it leads to frequent flagging of suspect periods. The PLS model (not shown) exhibits similar behaviour. Summary results are presented in tables 8 and 9.

Table 8: Number of hours or periods (in case of the moving averages) flagged by the models (only probable combinations considered) – Revised A13 motorway model

#*RMSE	PLS				OLS			
	1 hour	24 hour	1 week	4 weeks	1 hour	24 hour	1 week	4 weeks
1			58	16			93	36
2		73	23	5		65	12	1
3	190	17	1	0	116	10	2	0
4	47	0			22	0		
RMSE	7.68	3.89	2.49	1.89	11.92	5.98	3.50	2.55

Table 9: Number of hours flagged by one or both models for various indicators – Revised A13 motorway model

	Hour 4*RMSE	Day 4*RMSE	Day 3*RMSE	Week 2*RMSE
Both models flag	3	0	10	201
Only PLS flags	19	0	69	326
Only OLS flags	44	0	65	309
No flag	14642	14729	14585	13925
% agreement	5%	-	7%	24%

Though tables 7 and 9 cannot be compared directly as in the second case the period with problem data was left out, the measure of agreement between the models can be compared. If the period with the obvious deviations - on which the models agree - is left out the measure of agreement drops.

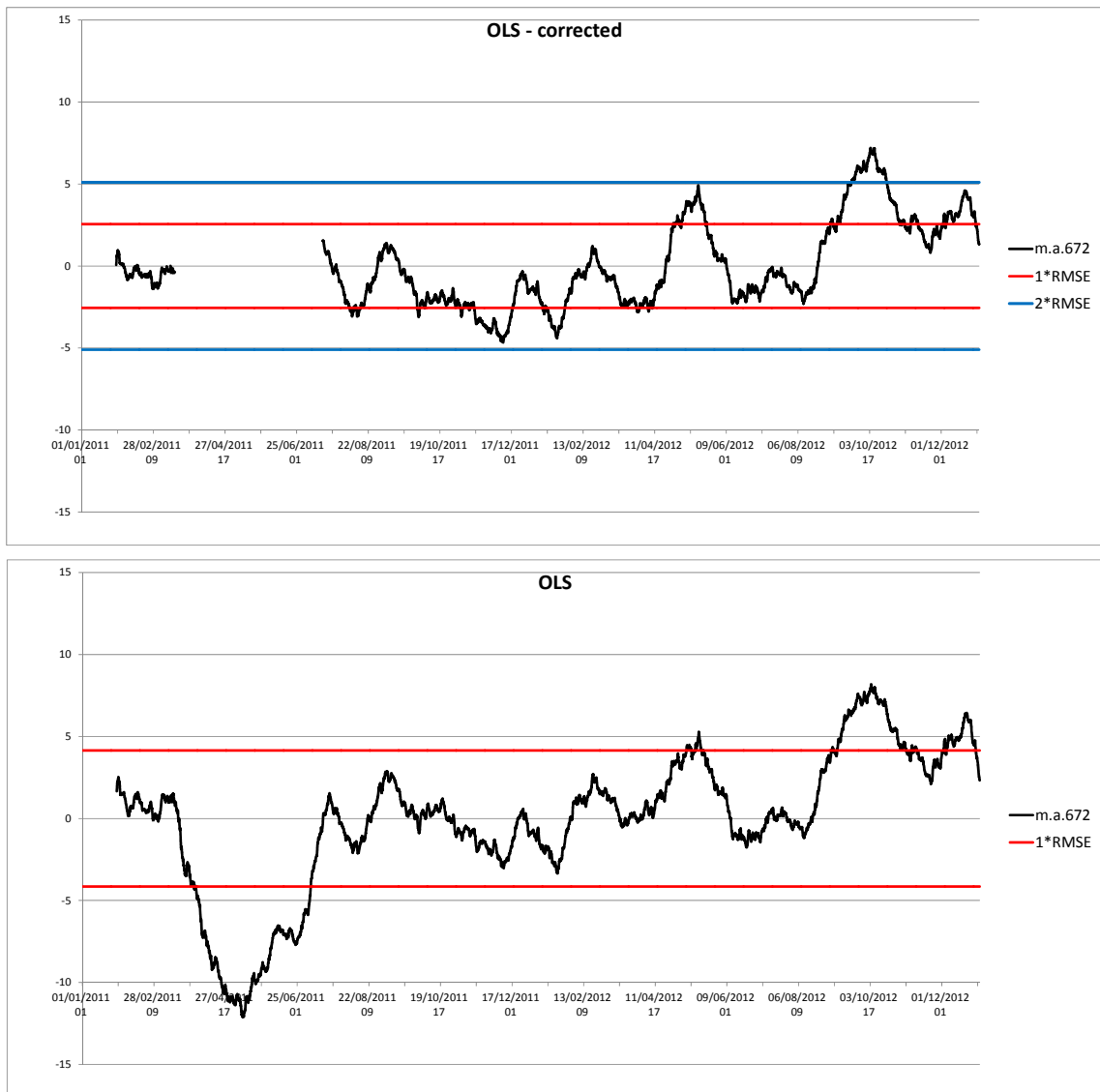


Figure 7: Two OLS models compared: model based on dataset with (bottom) and without (top) a period with the suspect measurements.

3.4 Stability of models over time

In the previous section data were analysed by models that were developed on the same data set. In reality one will be looking forward (new observations as they hourly arrive) with a model developed on a passed period. To examine if results obtained are sensitive to the time lag between the period when the models were developed and the moment of screening data, the models were applied to different periods in the past. Figure 8 shows an OLS model (4 week moving average) developed on the period 2011-2012 applied to the period 2003-2013. This was done for the Schiedam urban background station as this station didn't have issues in the period on which the model was developed. The red dots show moments when a monitor was changed either due to perceived problems or because of regular maintenance. The results show that the model is remarkably stable over a 10 year period. The graph shows that monitor changes often coincide with changes in the graph (note that not all maintenance actions are marked), implying that the larger changes in the graph indeed signify events. It also flags a few periods with potential issues.

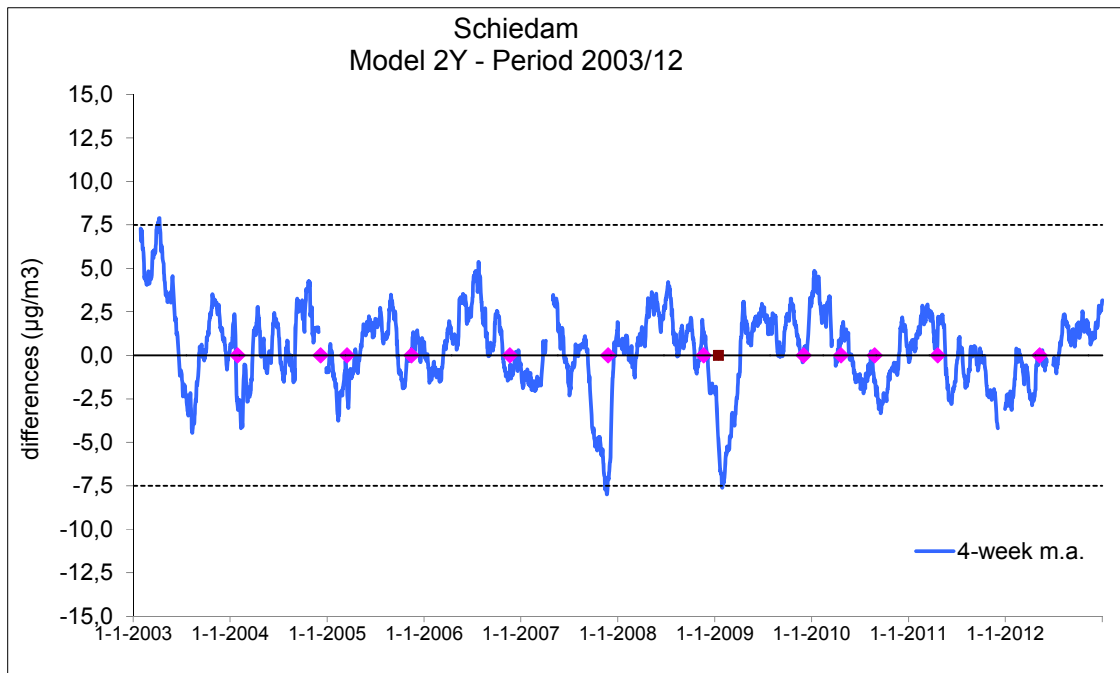


Figure 8. Ten year time series of a 4-week moving average based on a two year model – Schiedam urban background.

Though figure 8. looks quite stable over time the performance of the prediction model can be evaluated more formally by looking at the mean and the RMSE in each year. The mean is 0 for the period on which the model was developed and can be expected to deviate more as one moves away from that period. Similarly RMSE will be smallest for the period on which the model was developed and might go up as one moves away from that period. All this assuming that there are no issues in the period examined (clearly not the case in this example) that could cause a year to deviate from the overall trend. Figure 9 shows the evolution of both the mean and the RMSE.

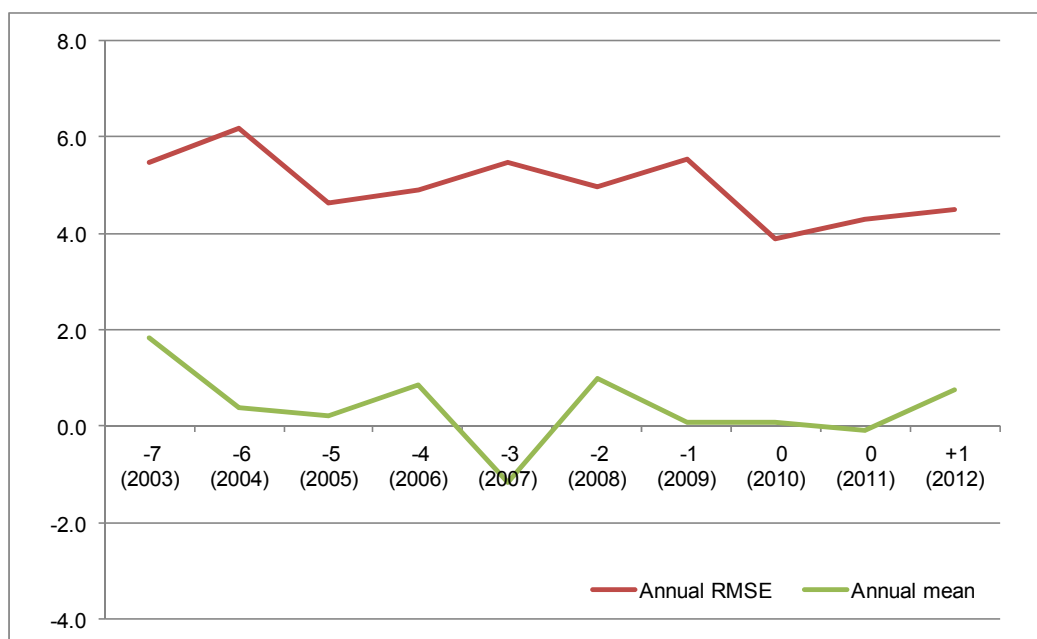


Figure 9. Evolution of mean and RMSE as a function of the years since the model was developed (OLS model developed on 2010-2011 daily observations – 4 week moving average).

Drift from 0 makes automatic screening less accurate and an increase of RMSE makes the models less sensitive to issues. In particular monitoring stations that are strongly influenced by a single source – such as a traffic station – that is subject to policy to reduce its impact, need periodic updating.

The results suggest that annual or biannual updates of the models will suffice to assure meaningful additional information from this screening method.

3.5 Testing for discontinuities after maintenance

Apart from the automatic flagging as suggested by the methods demonstrated so far, one could implement a test for a discontinuity after each maintenance event. Carslaw and Carslaw (2007) review various methods for detecting and timing unknown change points in trends and apply a method based on F-statistics (Zeileis et al., 2003) to formally test for changes. In this report we simply want to assess if a known event might have led to a change and hence we adopt a more simple approach by simply comparing the mean difference before and after maintenance. This test might rapidly detect potential errors but is unlikely to identify slow drift.

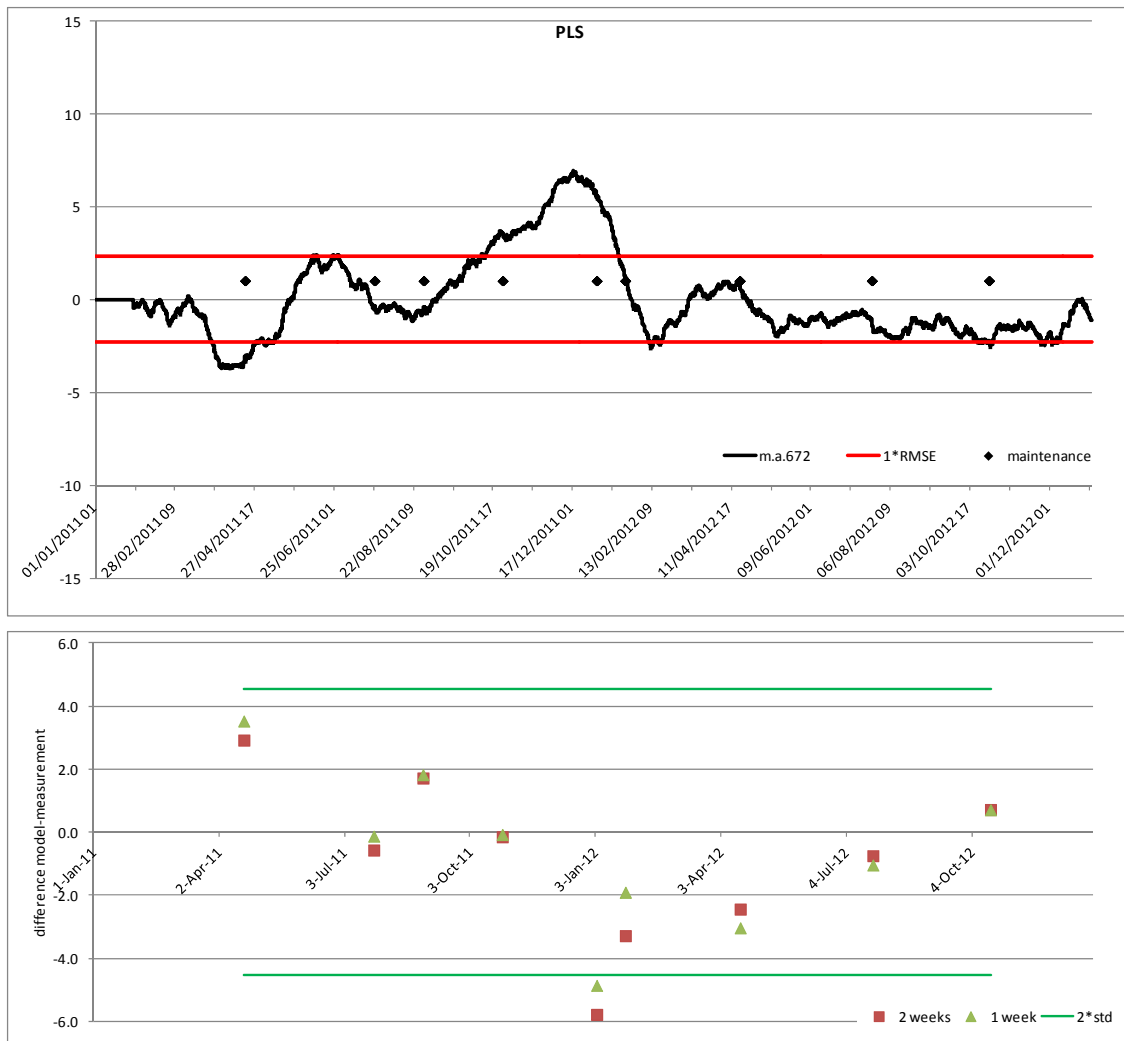


Figure 10: Top: monthly moving average Schieddamsevest - maintenance instances included. Bottom: average hourly *difference* before and after maintenance (average of 4 weeks before minus the average of either 2 or 1 week after the maintenance).

Figure 10 shows the results of a test whereby the hourly average differences between model and measurement, before and after maintenance are compared. Before the intervention a 4-weekly average is determined and this is compared to the average of a one or two week period after the maintenance. The results show that all but one instances of maintenance the differences before and after maintenance are less than 2 standard deviations. This is what one hopes and expects: the results before and after maintenance are sufficiently consistent. The only exception is not the start of the period of the slow upward drift, but the almost instant return to normal after the drift issue was corrected. This result suggests that this test can capture - within one to two weeks – a rapid change as a result of maintenance.

A similar test was done for the motorway station. See figure 11. note that this time there are two bands to indicate the 2 standard deviation range: one based on the period including the erroneous data and one without them. If the methods for the early detection of problems, as suggested here, are employed, the narrow bands are more likely too apply. The results show that both the rapid drop and the return to normal cause differences larger than 2 standard deviations. The test on the impact of maintenance would have identified the problem approximately two weeks before the 1*RMSE flagging of the 4-week moving average.

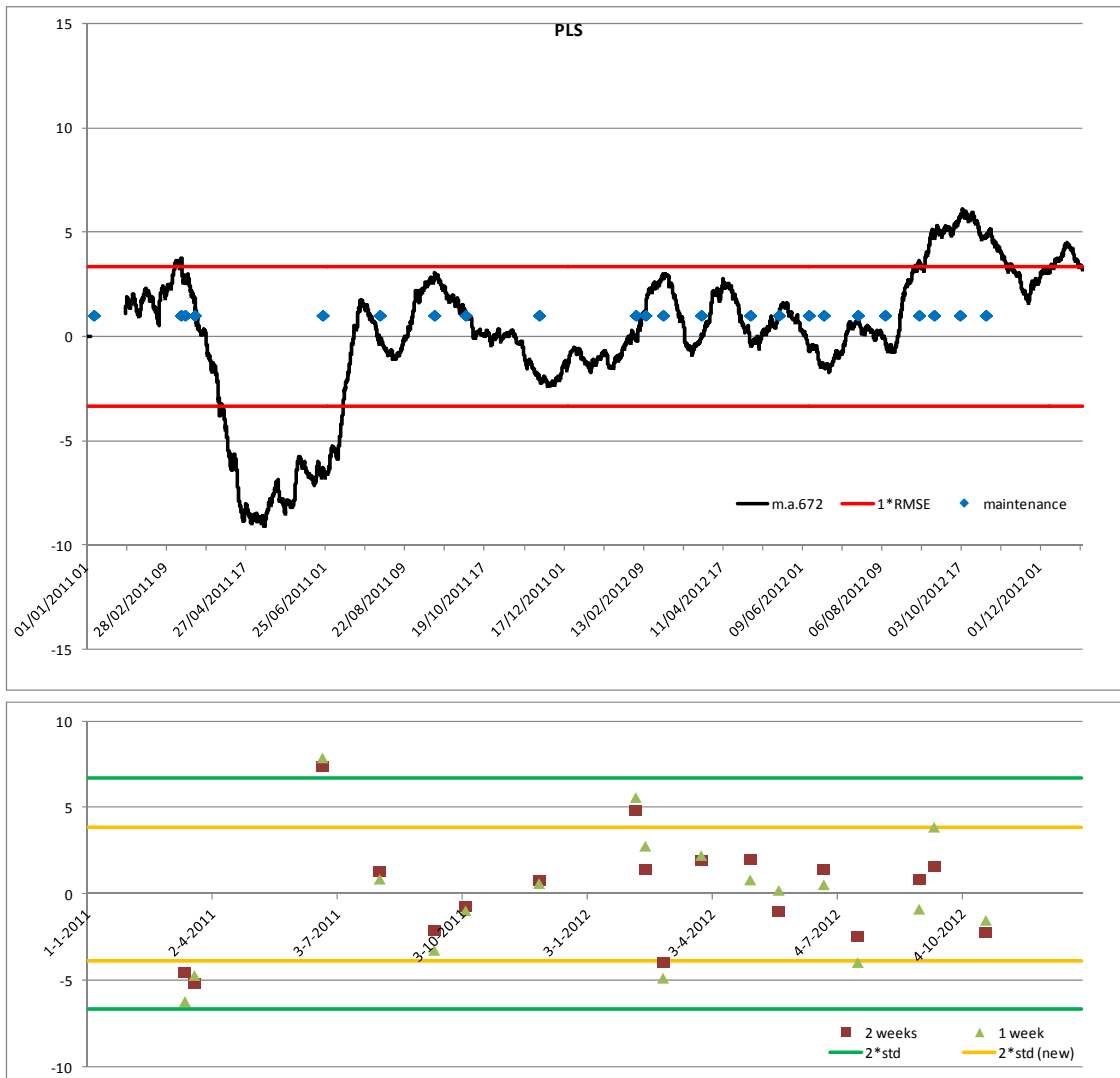


Figure 11: Top: monthly moving average A13 motorway - maintenance instances included. Bottom: average hourly *difference* before and after maintenance (average of 4 weeks before minus the average of either 2 or 1 week after the maintenance).

3.6 General applicability of the model approaches

The two cases described here deal with models to estimate measured values on nearby stations. As can be seen from figure 1, most of the stations are within 10 – 25 km distance of each other and there are a good number of background stations available to calculate an average background (the approach taken in the OLS models). Here we assess the potential of the method in an area where the monitoring stations are less dense.

Nguyen et al (2012) showed that monitoring sites in the Netherlands, depending on their type, exhibit fairly similar behaviour. This is exploited by using three monitoring stations in different parts of the Netherlands to make a model for a monitoring station in the study region. The stations used are at a distance > 50 km from the station for which a prediction is done and the distances between the three stations are also 50 -100km (or more). The rural background station is modelled using an average based on five other rural background stations. For the urban background and traffic stations an average background based on three urban background stations was used.

The results (OLS approach) are shown in table 10. As expected, the model performance is less than in the previous cases. The best performing model is for the rural background. That is the least demanding situation as the monitoring site is mainly subject to large scale influences that resemble those in the rest of the country. The performance goes down as the local influence becomes stronger. For reference the A13 motorway station from section 3.2 is included as well. The A13 column in table 10 can be compared to the A13 OLS model (2nd column) in table 5.⁹

Table 10: OLS modelling approach: regression coefficients using distant monitoring sites for model development.

	Westmaas - Rural background	The Hague - urban background	The Hague - Roadside	A13 Overschie - Motorway
Background concentration	0,91	1,03	1,1	0,94
Weekday (W)	2,94			7,26
Wind direction WR1	2,44	-4,71	-5,44	-17,91
Wind direction WR2	3,24	6,09	1,89	-10,40
Wind direction WR3	-0,68	14,14	6,75	9,62
Wind direction WR4	-1,90	5,75	6,22	11,46
Wind direction WR5	-2,63	-6,55	-0,94	9,58
Wind direction WR6	6,57	-11,04	-7,59	-8,76
Temperature	-0,38	-0,25	0,20	-0,17
Precipitation	-0,33			
(Windspeed+1) ⁻¹	33,00	29,12	31,42	28,81
Intercept	2,13	1,75	5,42	11,02
RMSE	9,44	13,99	14,00	14,93
R2_adjusted	0,63	0,59	0,55	0,54

Figure 12 compares the distant and the original model for the A13 motorway case. General behaviour is fairly similar though there are substantial differences in absolute terms. The model could be suitable for visual inspection of broad trends. Whether these models are suitable for automatic flagging and certainly at shorter averaging times is questionable. On the other hand one has to bear in mind that this is a rather extreme case where the stations that make up the average background are located in different parts of the country sometimes over 100 km apart,

⁹ The PLS model was also re-run using only stations from outside the study area (see fig. 1). As expected the model performance is somewhat reduced: R2-adjusted drops from 0.81 to 0.69 and RMSE increases from 9.4 to 12 (see also figure 12b).

likely facing different meteo conditions, etc. Graphs for the other locations are shown in Annex A3.

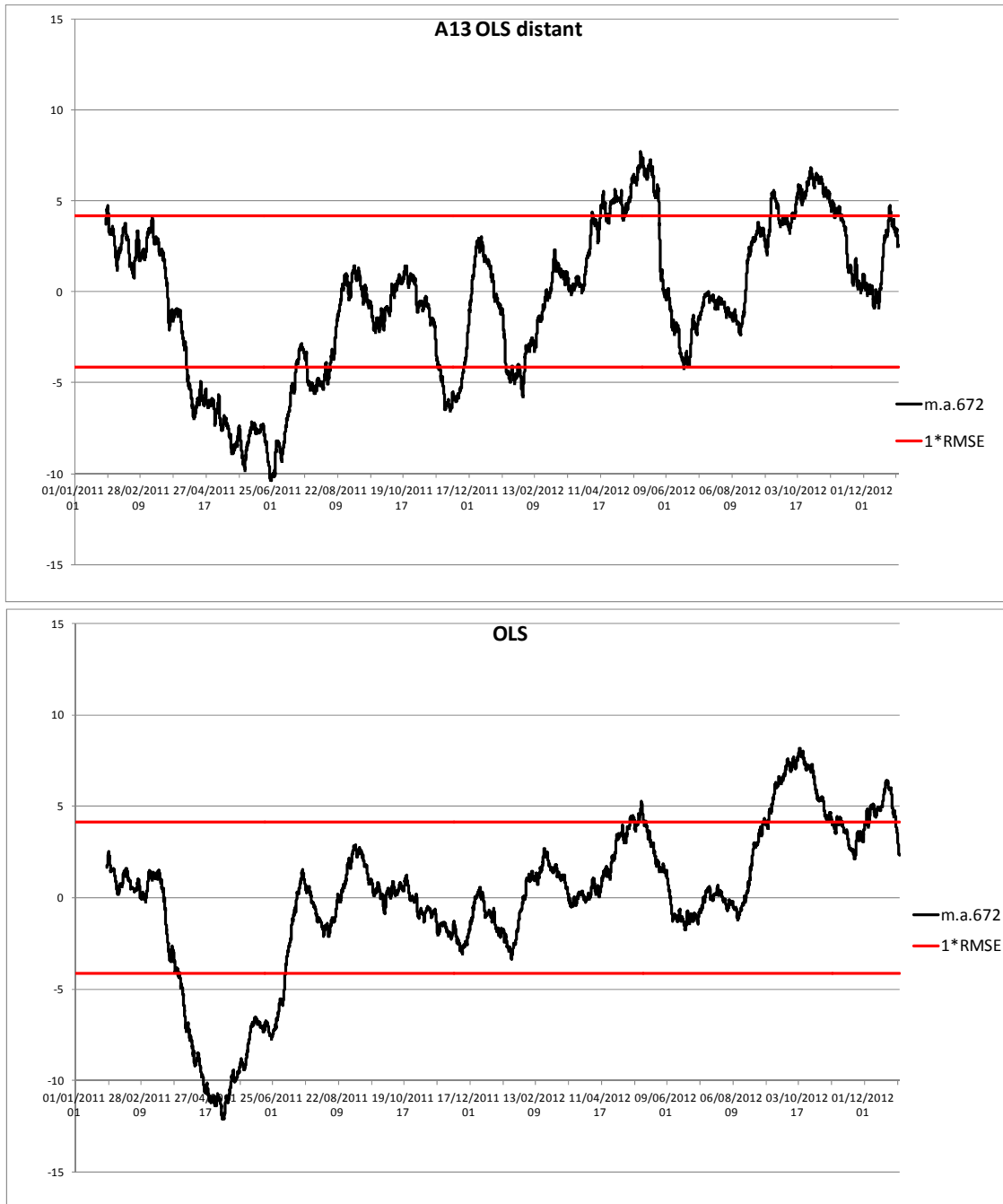


Figure 12a: Top: 4-weekly moving average A13 motorway – model based on three background locations far apart. Bottom: original model using five background locations in the same area (distance ≤ 25 km).

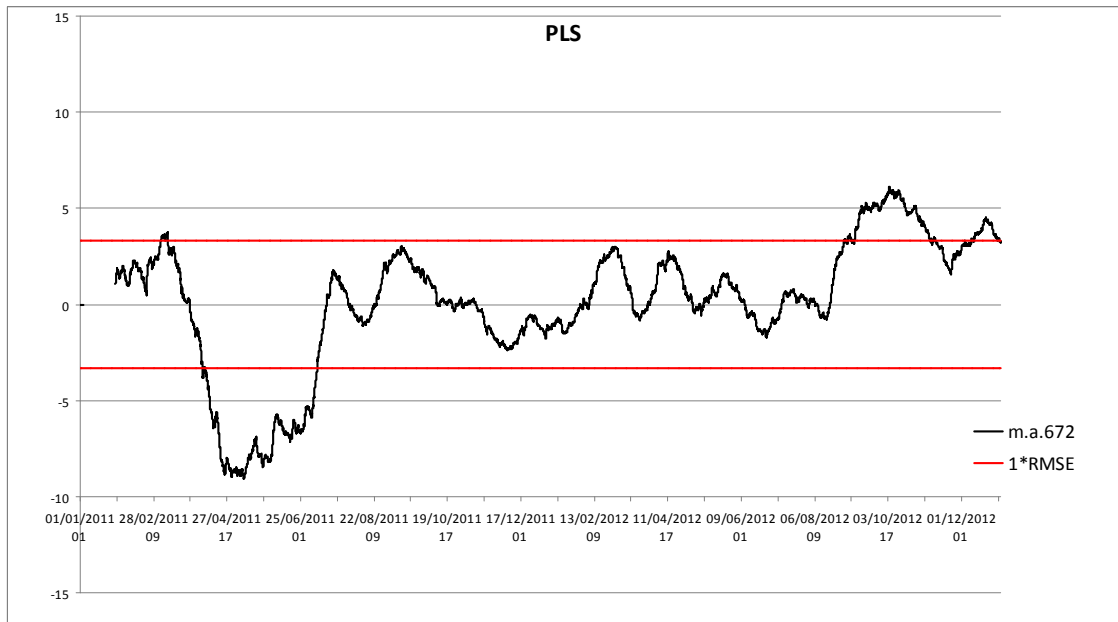
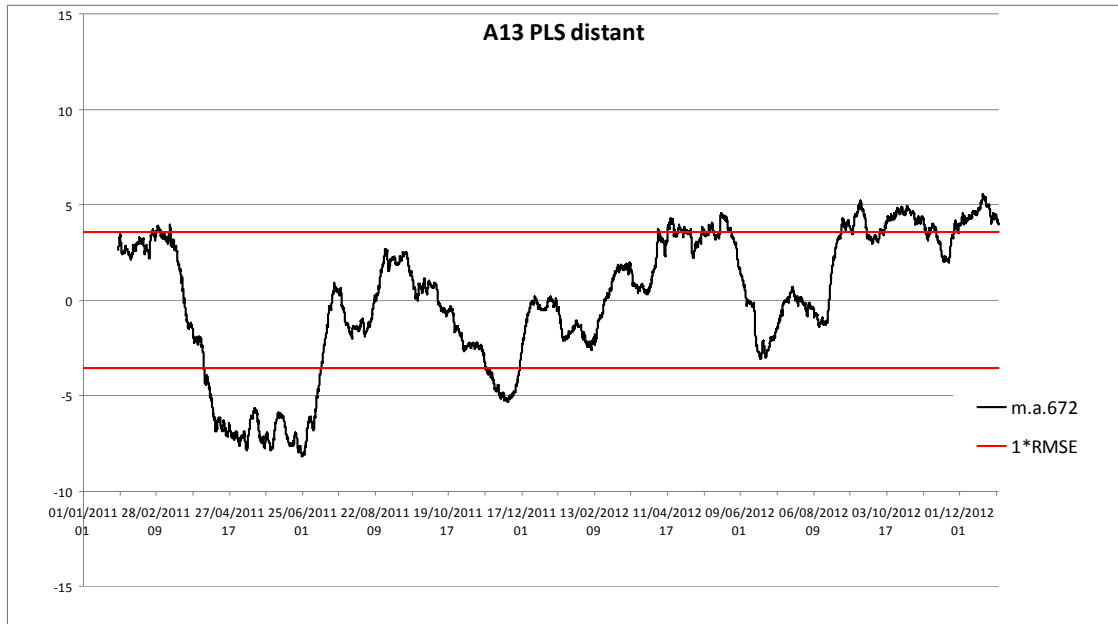


Figure 12b: Top: 4-weekly moving average A13 motorway – PLS model based on monitoring locations outside the study area. Bottom: original PLS model using nearby background locations.

4 Discussion and conclusions

4.1 Introduction

In chapter 1 we suggested that measurement data could be predicted (based on other measurements) and that these predictions can be used to make inferences on the probability of the actual measurement and hence provide a basis for the detection of anomalies. These anomalies – if serious enough - could imply:

- measurement errors
- changes in the way sources influence a monitoring site since the moment the prediction model was developed.

In both cases the flagged measurements would need additional attention during ordinary validation procedures. An example of the first case was shown in chapter 3. Figure 13 illustrates the second case. Most reversals of trends can be linked to maintenance events but in addition two periods of road works leading to reduced traffic seem to cause trend changes (local minima) in the difference between measured and predicted values.

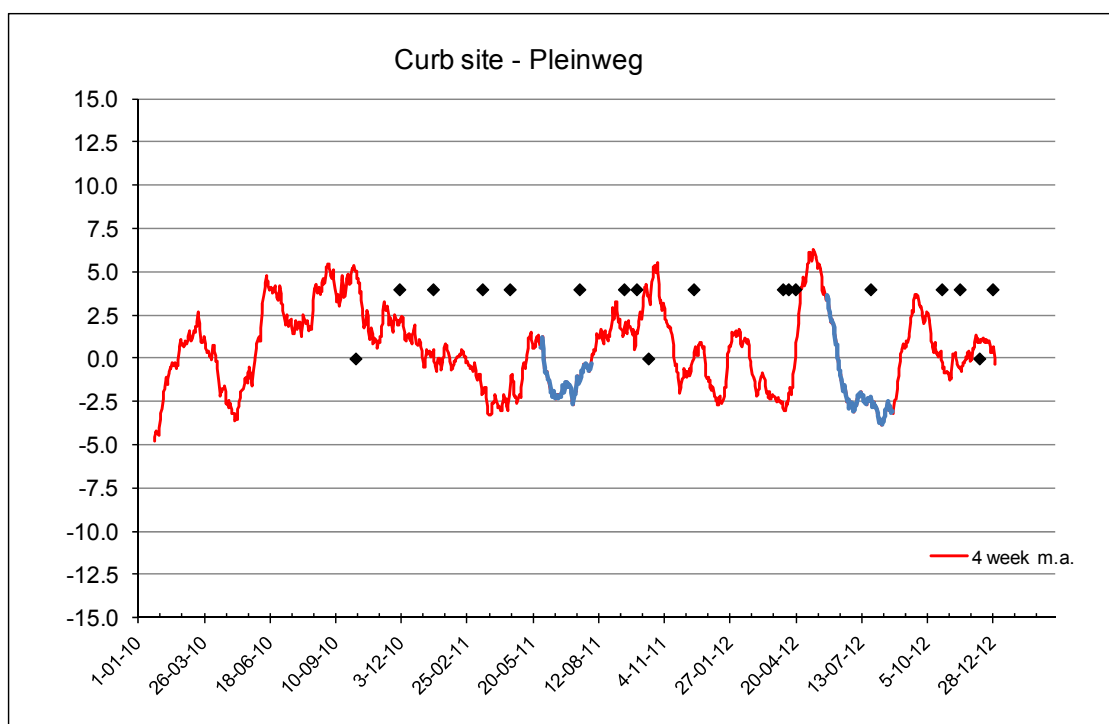


Figure 13: Curb site monitoring station and OLS modelling approach. The graph shows the changes of monitoring equipment (dot on the x-axes); other registered maintenance (other dots), and two periods with road works (blue line) upstream of the monitoring site reducing the traffic numbers.

In a recent application of the model at the Ridderkerk motorway site a downward trend was observed that was not completely rectified by maintenance. As the moving average remained too low it was decided to place an additional independent NO_x monitor to rule out any malfunctioning of the monitoring equipment. Rather than discovering strange behaviour in retrospect the model allowed us to perform additional tests during the seemingly deviating measurements. The research for this station now concentrates on a potential change in traffic numbers and data were requested from the motorway authority to further clarify the observed concentration drop.

The results in chapter 3 show that making measurement predictions provides useful additional information for the validation of measurements. The methods studied here revealed insights that previously went unnoticed, or were noticed but considered indecisive in the reference frame commonly used for validation. The methods proposed here provide a more formal reference context by quantifying the degree of deviation (in terms of RMSE). This could lead to a different judgement. In this chapter we will discuss the interpretation of the obtained results.

4.2 Validation support and early warning - general concepts

Measurements are surrounded by elaborate quality control systems assuring that measurement uncertainty remains within specified limits. This implies that not every deviation is immediately corrected. This would lead to a serious loss of observations and unrealistically frequent field visits to replace or adjust monitoring equipment. A certain margin of error is accepted (e.g. as in Shewhart quality control cards). Consequently some kind of scatter and/or a saw tooth pattern within the margins of tolerance is expected and no reason for action. Figure 14 shows the graph of an urban background station. It is a twin station of the curb site station shown in figure 13. The two are recently started monitoring sites, located within less than a km distance and serve as a pair to study the traffic contribution to air pollution.

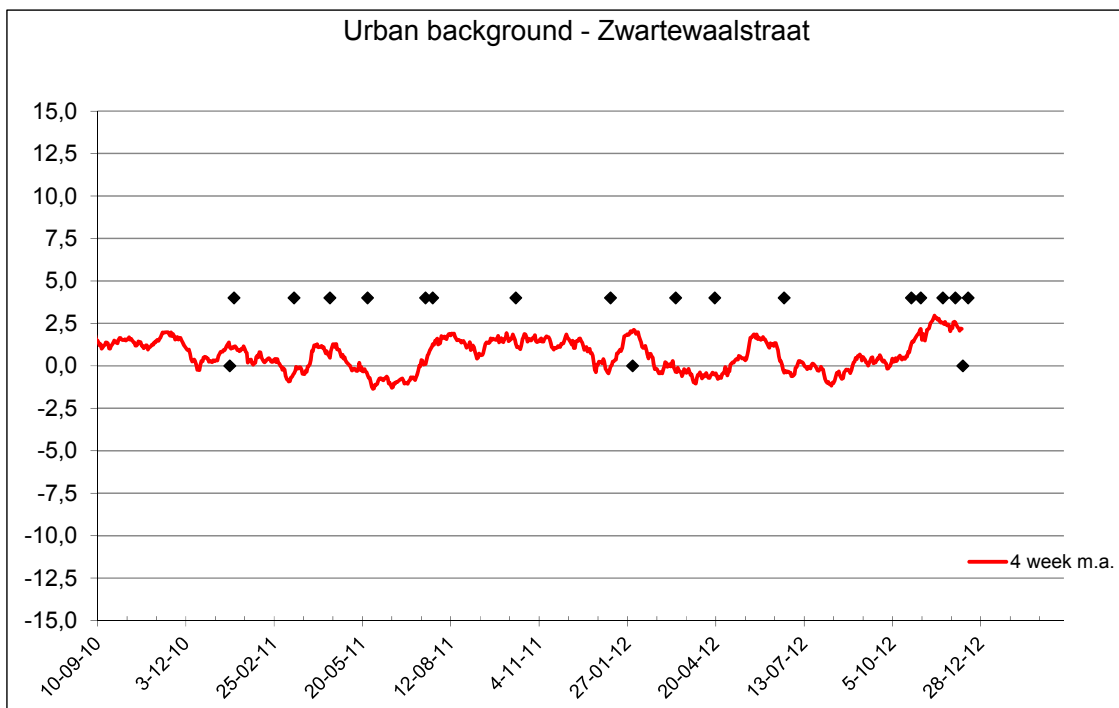


Figure 14: Background monitoring station and OLS modelling approach. The graph shows the changes of monitoring equipment (dot on the x-axes) and other registered maintenance (other dots).

The two graphs show that the variation differs between the roadside and the background monitoring site. In Chapter 3 we also saw that the motorway station had a higher RMSE than the background stations. While interpreting results this should be taken into account. In both graphs it is also visible that maintenance events increase in number during periods with larger deviations. Apparently suspicious behaviour was noticed during regular operations leading to additional inspections/calibrations.

The graph also shows that maintenance often coincides with changes in trends (note that there is some lag between a maintenance event and its appearance in the 4-week moving average). This is also very evident from figures 10 and 11 in section 3.5: many changes in the

trend can be attributed to maintenance events. Note that even small changes in trend are noticeable. Apparently the models are sensitive enough to capture these interventions if they lead to changes in the way the monitoring equipment behaves (calibrations, replacements of malfunctioning equipment, etc.),

Not every change has a meaning in terms of quality control and part of the saw tooth pattern is just a sign of regular operations. As the EN ISO 14211 (2012) indicates that NO_x measurements of the reference gas are to be guaranteed within 5% one could say that scatter up to 2 µg/m³ (5% at the limit value) is just a consequence of normal operations.

Another issue arises if there is a small gradual trend in one direction. Even if the deviations remain small in absolute terms this could be a sign that something is not working properly. For an operational system one would want to avoid this and rather strict criteria could be employed for the early detection of these trends. This could lead to additional onsite verification and/or maintenance. On the other hand, if screening as presented here is used in retrospect, it makes less sense to employ very strict criteria: why revalidate a lot of data that show, in absolute terms, only minor deviations? One could decide to only revalidate those hours that substantially exceed thresholds.

The results in Chapter 3 showed that the detection of broad trends is possible with both types of prediction models and that even biased models – due to questionable measurements in the dataset used for model development – are quite capable of detecting them. However, section 3.3 also showed that the thresholds used for decision making do strongly depend on the quality of the data used for model specification. Before deciding on a threshold the model should be run and if periods with substantial deviations do occur, the model should be rerun on a cleaned dataset. For operational purposes the model should be based on the best possible data.

So far thresholds for flagging were expressed in terms of levels of RMSE. In practice fixed numbers are easier to implement and the RMSE concept can be used as guidance for selecting an appropriate number. The exact level will depend on practical considerations where a feasible balance between the early detection of potential issues and the additional burden of hours to be investigated has to be found.

4.3 Interpretation of results

4.3.1 Model performance

In chapter 3 two modelling approaches were tested and compared. A simple multiple regression model (OLS) and a principal components based approach (PLS). The modelling approaches were different, the way the variables were used and the number of variables used were not identical. In a statistical sense the results should therefore not be seen as a competition between the two modelling techniques. However, the results show that the PLS modelling approach did perform somewhat better than the OLS approach. This is mainly relevant when judging hourly or daily results. In a real-time system (e.g. screening data before they are put online) the better model obviously has an advantage. If the averaging time increases and the methods are used to detect broader patterns both approaches produce similar results. In that case the OLS models have a major advantage in the sense that they are very transparent in what is happening. For example: suppose a deviation occurs on a public holiday. The regression coefficients (as in table 5) could quickly reveal that for the motorway station a difference of 5.4 µg/m³ would be expected as the model thinks that it is a working day. Similarly, the influence of the wind direction¹⁰ can easily be traced.

The method discussed in this document is based on a concept used by Carslaw and Carslaw (2007). It compares a measurement with an estimate of the measurement. They use

¹⁰ In this study we used 6 fixed wind sectors for the OLS. If a monitoring site is specifically influenced by one or more identifiable sources these sectors could be adapted to get a better OLS model.

Generalised Additive Modelling (GAM) to forecast daily averaged NO₂ measurements based on the NO_x concentration, background O₃ and NO₂ concentration and a wind vector. The method was applied to 20 roadside stations in London. They obtain r² values between 0.74 and 0.97. Results for r² obtained here in chapter 3 range from 0.64 (or 0.71) to 0.88 depending on the modelling approach and whether or not a period of problem data were excluded from model development. The fact that the r² are somewhat lower in this study may have several reasons. Firstly, hourly data are used instead of daily data leading to higher variability in this study. Secondly, Carslaw and Carslaw include total NO_x measured at the site where NO₂ is estimated in their equation. Lastly, they explain explicitly that the data used to build the model should be free of bias. In this case no a priori screening of the data was done in order to mimic real time application of testing unvalidated data.

4.3.2 Hourly and daily differences between measurements and predictions

The results in chapter 3 showed that both modelling approaches flag a good number of hours as suspicious in addition to the hours already rejected by the validators. So an additional statistical screening of the measurements (as suggested in this document) seems to provide useful further guidance for validation. However there are certain issues as well:

- The models partly flag different hours;
- Depending on the threshold many additional hours are flagged for further investigation;
- Sometimes hours that were rejected on technical grounds were accepted by the prediction models.

Hourly data: Since the agreement between the prediction models and or the model and the validator is only partial and since it is impossible to know whose judgement is right or wrong it is recommended to only flag seriously deviating hours/days. For operational purposes hours with a flag could be left out of real-time presentations and further scrutinised during validation. Flagging individual hours or days is sensitive to model bias so particularly for this purpose the best possible dataset should be used for model development.

If a filter is desired to support real-time publication of hourly values, 30 µg/m³ is recommended. This would filter 85 and 50 hours at the background and motorway station respectively in a two year period. This amounts to approximately 0.5% of the data stream.

Daily data: Based on the results daily deviations of 12.5 µg/m³ are flagged. This corresponds to approximately 3*RMSE (depending on the site and the quality of the data used to develop a model). This results to 30 and 3 hours in the background and the corrected motorway datasets respectively in a two year period. For validation guidance the daily variable is suitable.

The exact threshold should be based on practical experience and the amount of validation support one wants to receive. One could argue that if the thresholds are set in a such a way that less than 1% of the (unvalidated) data are flagged it might not be worthwhile maintaining the system.

4.3.3 Detecting trends

For the detection of trends the longer averaging times are more suitable and the models are less sensitive to bias. The 4-weekly moving average is easiest to interpret but might delay the detection of suspicious data. Here we compare the 1 week and 4-weekly results.

The urban background case in section 3.1 showed that an analysis based on a one week moving average at 1*RMSE would lead to early detection but also to several periods being flagged. A 2*RMSE threshold would delay the detection of the drift. The combination of a 4-weekly moving average and 1*RMSE threshold gave the same time of detection as the 1 week/1*RMSE combination. However, the signal is cleaner and easier to interpret and the number of periods flagged is less. Hence the 4-weekly moving average is the preferred indicator.

In case of the motorway station with the rapid drop in concentration after maintenance the situation is different. The detection using the 4-weekly moving average is notably later than the weekly criterion. However, also in this case the 4-weekly moving average signal is much easier to interpret than the weekly signal. See figure 15. If really the first passage of the 1 or 2*RMSE threshold is interpreted (left most vertical blue line), the weekly criterion would be faster. However, the weekly curve bounces back to normal quite quickly. This could be interpreted as a false alarm and if the decision to further investigate is only made after the second passage there would be no advantage in using the weekly curve over the 4-weekly curve.

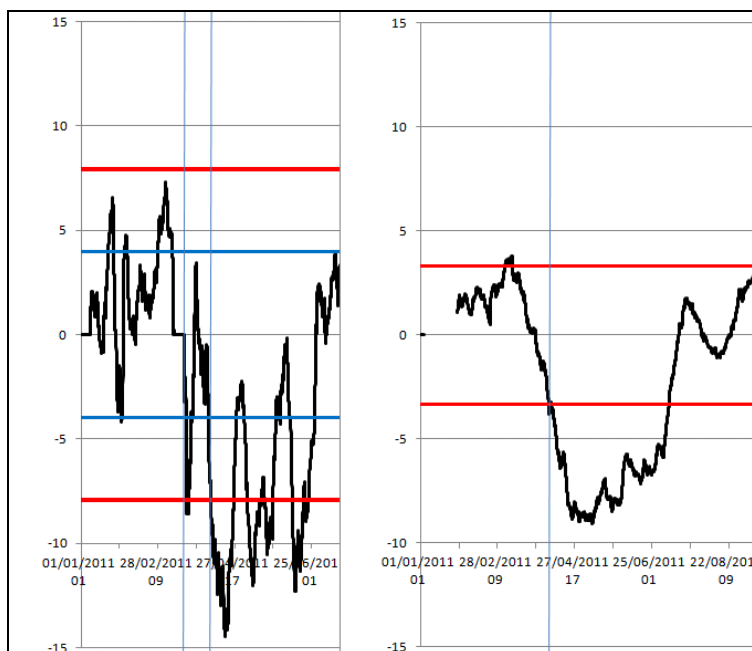


Figure 15: PLS models with a weekly (left) and a 4-weekly (right) moving average. The timing of decisions depends on the exceedence of the thresholds: left 1 or 2*RMSE (blue and red lines respectively); and right 1*RMSE (red line).

Four weekly moving average: for automatic warnings a threshold of $3 \mu\text{g}/\text{m}^3$ can be used. This is approximately 1 to 2*RMSE depending on the modelling approach and the correctness of the data used to develop the models.

Also in this case the exact threshold should be based on practical experience and perhaps set differently for urban background and traffic monitoring sites. After all if the applied method leads to early detection of issues and less noise in the data, previously functional thresholds might lose their effectiveness (absolute threshold never being exceeded, RMSE based threshold becoming too strict).

For a rapid detection of quick changes after major maintenance such as replacements of a monitor, a gas bottle etc.) a t-test on the average before and after the maintenance is probably the best way. See the results in section 3.5.

4.4 Generalisation of results

The methods developed in this report predict measurements using a number of meteo and day of the week variables as well simultaneous measurements at other locations and succeed (surprisingly) well in doing so. The study area is well equipped with monitoring stations and

stations used in building the models are located within a relatively short distance from each other. In section 3.6 (and Annex A.3) results were shown from the opposite situation where sites far apart were used as input for the OLS model used for the measurement estimates. The results show that the general concept still works but the results is less robust if one has to rely on few monitoring stations.

4.5 Conclusions

In this study it was shown that models can be made to predict measurements. The models are sufficiently accurate to act as a first check on the likeliness of the measurement observations. Analysing the differences between measurements and predictions is a useful way of obtaining additional information for the validation of the air quality measurements. Two monitoring sites were studied in depth and examples of an additional five monitoring sites are given. All in all the method was applied at 10 sites in the study area. The concept of estimating measurements to check the behaviour over time of these measurements, was also demonstrated for 20 traffic monitoring sites in London (Carslaw and Carslaw, 2007).

It was shown that the statistical screening of measurement data by using predictions based on simultaneous measurements at other locations is a useful addition to regular validation.

Different parameters are needed for different purposes:

- A daily moving average of the difference between measured and predicted observations can be used to flag for outlier measurements needing additional attention during validation.
- Hourly differences can be used to detect potential deviations in real-time applications.
- 4-weekly moving averages can be used to detect cases of drift that are either due to changing circumstances or malfunctioning of the measurement set-up.

The most accurate models are derived by PLS using the information from both meteo variables and other air quality measurement stations. Accurate models are particularly important for real time application when one has to judge highly variable hourly values. If the methods are mainly used as validation guidance and if one focuses on daily and 4-weekly averages the OLS models will do. These models are slightly less accurate but easier to develop and maintain and transparent in their operation. To detect issues in trends even simple regression models developed on unvalidated data including trouble periods can be used.

Models can be updated biannually unless a known change in the emission sources influencing a station has occurred. The trend detection is rather insensitive to the model being accurate. If the models are used for the detection of outliers on an hourly or daily basis, the best possible models are needed.

The models in this case study were based on two years of monitoring data. This was intuitively assumed to be a reasonable compromise between the number of data needed and the sensitivity to exceptional meteo or deviating data in the data set. The use of one year of data is possible as well provided trouble data are excluded from model development (see for example Carslaw and Carslaw, 2007).

The usefulness of the tool depends on the quality of the prediction model. In densely monitored areas (as studied in this report) it is possible to develop good models. The further apart the monitoring stations are, the harder it becomes to develop adequate models. The study showed that models developed with monitoring stations up to 100 km apart might still provide useful complementary information for the validation process.

5 Literature

- Carslaw, David C. and Nicola Carslaw. 2007. Detecting and characterising small changes in urban nitrogen dioxide concentrations. *Atmospheric Environment* 41 (2007) 4723–4733
- CEN/CENELEC. 2005. General requirements for the competence of testing and calibration laboratories (ISO/IEC 17025:2005).
- Elshout, Sef van den. 2003. Betrouwbaarheid van discontinue luchtkwaliteitsmetingen. *ArenA/Het Dossier*. Jaargang 9, p143-146.
- EN 14211 (2012). Ambient air - Standard method for the measurement of the concentration of nitrogen dioxide and nitrogen monoxide by chemiluminescence.
- Mooibroek D., J. Vonk, G.J.M. Velders, T.L. Hafkenscheid, R. Hoogerbrugge. *PM_{2.5} Average Exposure Index 2009-2011 in the Netherlands*, RIVM report 680704022/2013.
- Nguyen, P.L., Stefess, G., de Jonge, D., Snijder, A., Hermans, P.M.J.A., van Loon, S., Hoogerbrugge, R. 2012. Evaluation of the representativeness of the Dutch air quality monitoring stations. RIVM report 680704021
- Zeileis, A., Kleiber, C., Kramer, W., Hornik, K., 2003. Testing and dating of structural changes in practice. *Computational Statistics and Data Analysis* 44, 109–123.

Annexes

A1 PLS modelling: additional trials using unvalidated data and inverse wind speed

Validated versus unvalidated data

The model described in this report was developed on validated data. In the validation process unvalidated data will be used. This aspect was examined with PLS modelling using the dataset Schiedamsevest (urban background). The unvalidated 2 years-dataset of Schiedamsevest contains 148 data which were rejected in the manually validation. Because the PLS tool can not deal with missing data, prior to modelling, missing data (no measurement available and data which were rejected) were filled by values which were estimated based on available data, using a simple formula that calculates missing value x_{ij} as the average of all stations at hour i adjusted with the ratio of the overall average to the average at site j according to:

$$x_{ij} = \frac{\sum_{k=1}^N x_{i,k(k \neq j)}}{\sum_{n=1}^{17544} \sum_{k=1}^N x_{n,k(k \neq j, n \neq i)}} * \sum_{n=1}^{17544} x_{n,j(n \neq i)}$$

$x_{i,j}$: estimated concentration of hour i at station j
 N : number of stations

This procedure was applied in all PLS models described in this report. It has to be remarked that if a large part of the dataset was missed, the above formula causes a bias in the filled up data. To avoid a bias, *average* instead of *sum* should be used.

In the table below the comparison between PLS models developed on validated and unvalidated data respectively, is shown.

Table A.1: Amount of hours flagged by one or both models for various indicators

	Hour 1*RMSE	Hour 2*RMSE	Hour 3*RMSE	Hour 4*RMSE
Both models flag	3731	766	232	85
PLS-validated dataset flag	328	72	15	3
PLS-unvalidated dataset flag	307	68	21	8
No flag	13178	16638	17276	17448
% agreement	85	85	87	89

We can conclude that the use of unvalidated data does not have significant impact on the results

Wind speed versus inverse wind speed

When the wind speed is close to zero the effect of wind speed is much more pronounced if inverse wind speed instead of the wind speed itself is used. This effect is examined by PLS model using the dataset Schiedamsevest. The table below shows the comparison between both models.

Table A.2: Amount of hours flagged by one or both models for various indicators

	Hour 1*RMSE	Hour 2*RMSE	Hour 3*RMSE	Hour 4*RMSE
Both models flag	3896	809	240	86
PLS-dataset with WS flag	163	29	7	2
PLS-dataset with inverse WS flag	133	24	9	2
No flag	13352	16682	17288	17454
% agreement	93	94	94	96

We can conclude that though the inverse wind speed has a theoretical advantage, over wind speed as such, the practical results are quite similar.

A2 Additional information model stability

OLS models were made for various stations in the Rotterdam area. The models were fitted on 2011-2012 data and were applied for the period 2003-2012. Contrary to the models in the main text, the models discussed here are based on daily averaged concentrations. For Schiedam the ten year results were shown in figure 8. In this annex a few more examples will be discussed.

Almost all graphs show increasing scatter the further one moves away from the data period on which the models were developed. This is something one could expect from any model extrapolation. There is a second reason why this is happening: over time the ISO standard that governs the measurements has changed. Previously (before 2010) deviations up to 10% in the periodic span checks were allowed before readjustment of the monitoring equipment took place. Currently this threshold is set at 5%. Climate control at the monitoring sites was also improved over the years.

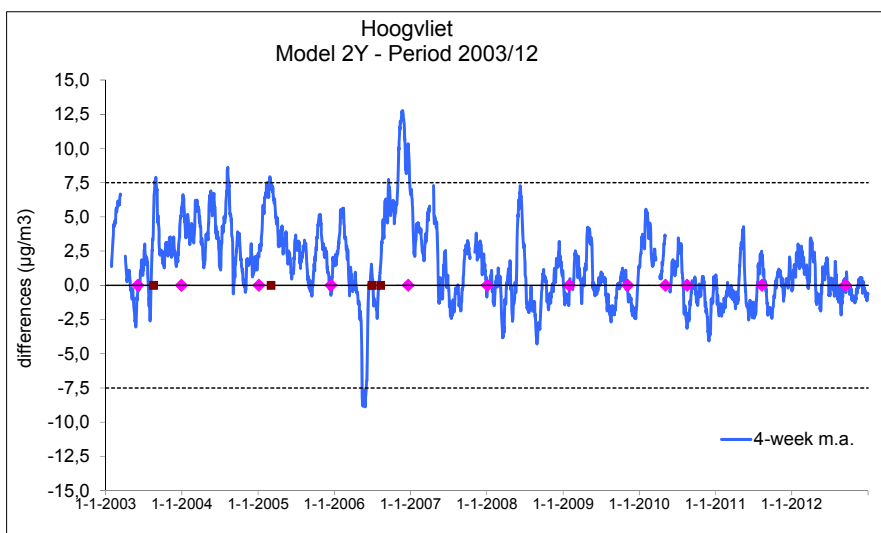


Figure A.1 Hoogvliet 10 year back extrapolation of the 2011-2012 model.

Background site Hoogvliet shows a quite stable graph over the years. Like Schiedam the 2-year model seems to be adequate over the full year period. Like in Schiedam, the historic data show two strange peaks, that were missed at the time during validation.

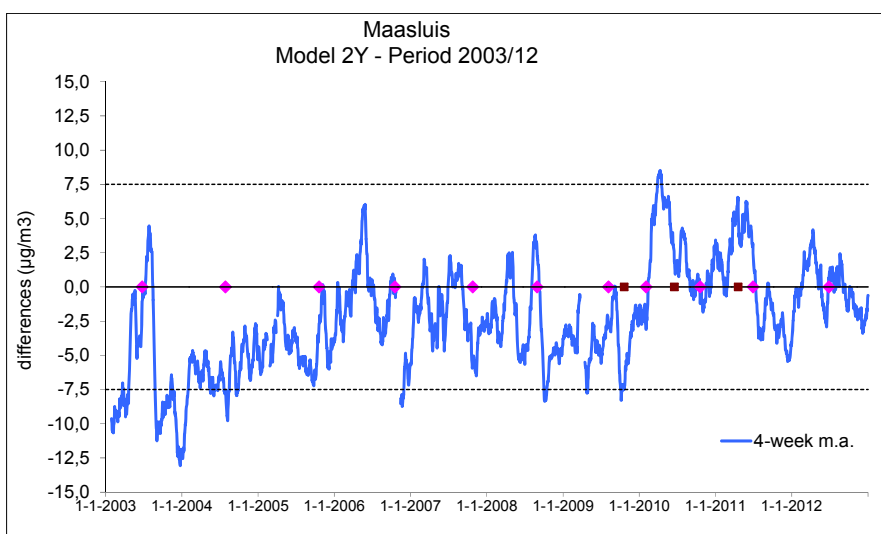


Figure A.2 Maasluis 10 year back extrapolation of the 2011-2012 model.

The Maassluis graph is an example of an urban background station that was not stable over time. In different periods construction work took place near the monitoring site and over the years the sources that influence the station have changed. The residential area is now more dense and closer to the monitoring station.

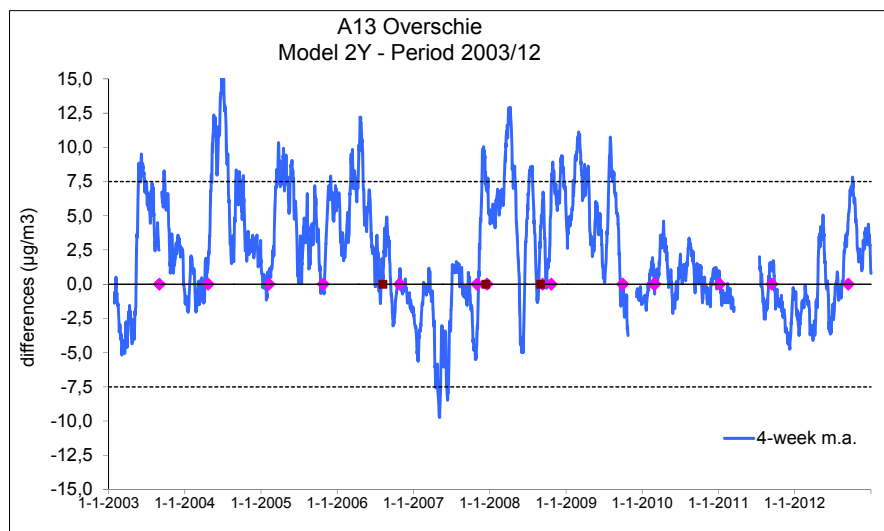


Figure A.3 A13 Overschie 10 year back extrapolation of the 2011-2012 model.

At motorway site A13 -Overschie the 2-year OLS model exhibits an upward trend if we go back in history. This can easily be understood. In essence the model describes how the motorway station behaves relative to the prevailing background concentrations and over the years the gap between traffic and background stations has been reduced. The transport sector has been subject to elaborate emission control legislation reducing the emissions per/km substantially. Secondly, road works north of the monitoring have widened the road and reduced the number of traffic jams for north-bound traffic. This has further reduced the concentrations at this site. In this case the model that provides the estimates would need an annual or biannual update.

Table A.3: OLS coefficients

	Dataset used: 2003-2004	Dataset used: 2011-2012
Background concentration	0.82	1.00
Weekday (W)	7.79	5.70
Wind direction WR1	-6.20	
Wind direction WR2		
Wind direction WR3	18.33	16.51
Wind direction WR4	21.11	20.67
Wind direction WR5	20.55	17.69
Wind direction WR6	9.01	9.87
Temperature	0.18	
Precipitation		
(Windspeed+1) ⁻¹	22.18	
Intercept	-3.91	-6.30
RMSE	9.564	7.451
R2_adjusted	0.74	0.82
Variables used	10	7

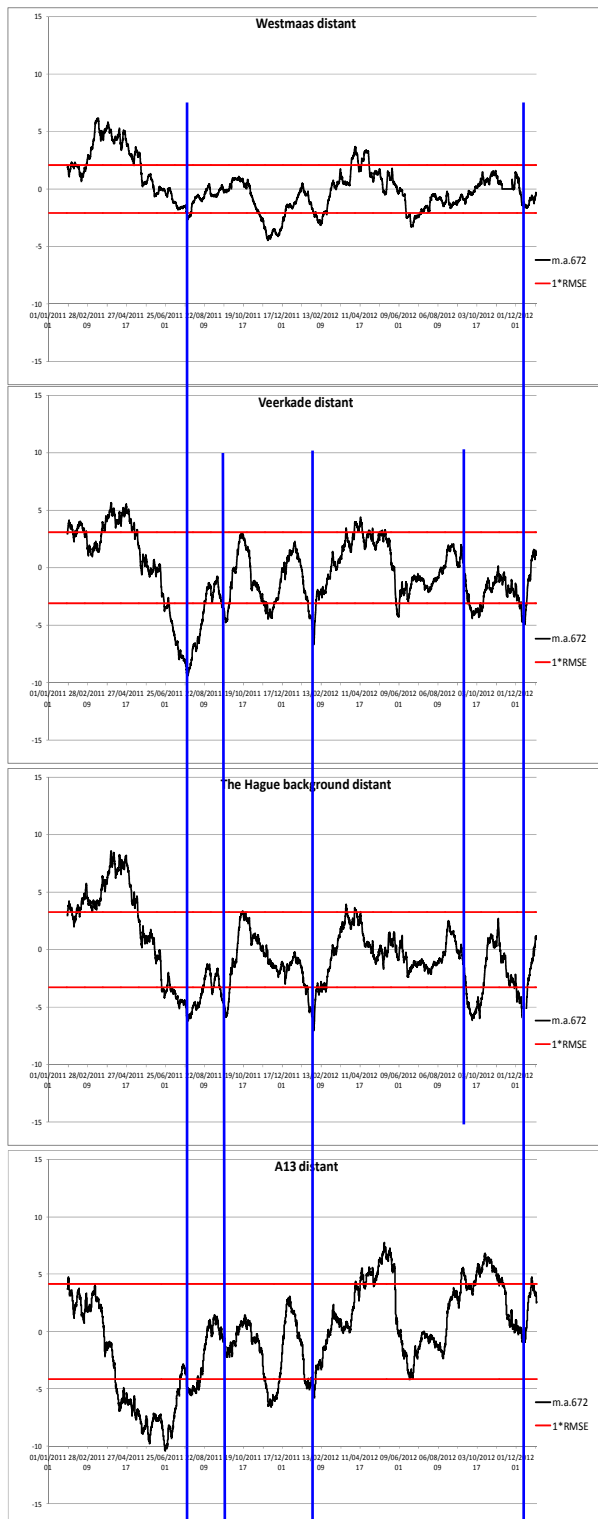
Comparing both models it is evident that the motorway contribution has changed. The additional working day increment has dropped by $2 \mu\text{g}/\text{m}^3$. This signifies that the difference between week and weekend that is due to the difference in traffic density was reduced. Also the contributions in the wind directions 3, 4 and 5 have gone down. This covers the compass angles from 120 to 300 degrees: the exact location of the motorway as can be seen from figure A.4 where the monitoring station is marked with the arrow. South-westerly winds dominate in the Netherlands so this marks a noticeable reduction.



Figure A.4: A13 Motorway and location of the monitoring site.

A3 Results using distant background stations (OLS models)

In section 3.6 the regression coefficients were shown for four cases where background station at major distance > 50 -100 km were used to built the estimation models.



The first graph is for a regional background monitoring station and the average background was derived from the sites (Posterholt, Eibergen, Wekerom, Balk and Kollumerwaard). The results appear quite straight forward.

The other graphs are urban situations. The background was derived from the sites (Heerlen, Nijmegen and Groningen). The second and the fourth graph are traffic influenced sites, the third is an urban background site. Graphs two and three are in The Hague, the last one is in Rotterdam (and was shown in section 3.6).

The reason to include it again is that the last three appear to share some spikes. Some even seem to coincide in all four graphs. If a spike occurs in all four it could hint at meteo differences between the study area and the places where the background concentrations were determined. A 4-week moving average is presented so it is unlikely but it can't be ruled out.

Alternatively something could be happening at the background stations (and indeed there were periods with missing values) that were used. It is an average of only three stations so it is less robust than the background used previously (based on four or five stations in the same area).

A local change mainly affecting The Hague seems to occur (maintenance on the same day in the same area?). See second blue line on the right.

Figure A.5: 4-weekly moving averages for four sites based on distant background stations

The left most blue line shows that several effects could coincide. At first glance the bottom graph (A13) doesn't seem to behave like the others. However, if one looks at the original graph (see figure 12) one can see that there was a strong upward movement of the moving average at the time if the local model is used. What appears as a small dip in the bottom graph of figure A.5 is in fact a major downward movement at the same moment as in the other graphs.

Since this is happening at all four locations it is not only a background influence. The deviation is rather big, so only meteo influence is unlikely. Maintenance on the same day has to be ruled out as all background stations are far apart. Simultaneous maintenance on the study stations is unlikely as one is managed by another organization and only two are conveniently close. Just after the dip in the graphs there are missing values in the stations used to provide the urban background. Perhaps something was going astray that needed maintenance (the period with the missing values). This could (partially) explain the dip but it is not the only explanation.

The variability in the circumstances that influence the shape of these graphs is considerably higher than that in the cases where all stations considered are relatively close to each other. The current report has not analysed what the maximum distance between monitoring sites could be. The preliminary tests shown here demonstrate that in these cases it is important to use sufficient monitoring stations to determine the average background (in case the OLS modelling approach is used) and that complementary information such as maintenance dates is definitely needed to interpret the graphs. In short, the method might still be useable, but relying on automatic screening is more difficult as more complementary information is needed.

DCMR
Milieudienst Rijnmond
Parallelweg 1
Postbus 843
3100 AV Schiedam
T 010 - 246 80 00
F 010 - 246 82 83
E info@dcmr.nl
I www.dcmr.nl
Twitter: @MilieuRijnmond

