ARCHIVES OF ENVIRONMENTAL PROTECTION
vol. 39 no. 4 pp. 123 - 144 2013
VERSITA
PL ISSN 2083-4772
DOI: 10.2478/aep-2013-0041


# DATA MINING SYSTEM FOR AIR QUALITY MONITORING NETWORKS

PIOTR CZECHOWSKI[1]*, ARTUR BADYDA[2], GRZEGORZ MAJEWSKI[3]

[1] Gdynia Maritime University, Information Systems Department,
Morska 83, 81-225 Gdynia
[2] Warsaw University of Technology, Environmental Engineering Faculty,
Nowowiejska 20, 00-653 Warsaw
[1] Warsaw University Of Life Sciences – SGGW, Department of Meteorology and Climatology
Nowoursynowska 166, 02-787 Warsaw
* Corresponding authors e-mail: oskar@am.gdynia.pl

**Keywords:** Air pollutants, data quality analysis, data mining, estimation, exploratory methods, harmonic, Principal Component Analysis, statistical tests for outliers, influential, leverage observations, analytical software system, programming, EDM, Winsorized mean, Cook, Mahalanobis distance, DFITS, COVRATIO.

**Abstract:** The use of quantitative methods, including stochastic and exploratory techniques in environmental studies does not seem to be sufficient in practical aspects. There is no comprehensive analytical system dedicated to this issue, as well as research regarding this subject. The aim of this study is to present the Eco Data Miner system, its idea, construction and implementation possibility to the existing environmental information systems. The methodological emphasis was placed on the one-dimensional data quality assessment issue in terms of using the proposed QAAH1 method – using harmonic model and robust estimators beside the classical tests of outlier values with their iterative expansions. The results received demonstrate both the complementarity of proposed classical methods solution as well as the fact that they allow for extending the range of applications significantly. The practical usefulness is also highly significant due to the high effectiveness and numerical efficiency as well as simplicity of using this new tool.

## INTRODUCTION

Study regarding Eco Data Miner (EDM) project started back in the late 1990s. Professor Jerzy Trapp [20] was an academic supervisor of the project, accompanied by his team. In the initial phase, various software solutions were used. In the course of research being developed, the amount of received data has forced the researches to design more complicated systems, dedicated particularly to the research being done. These systems did not create an integral environment, realizing narrow/specific tasks, sometimes they tended to double one another. At the beginning of the present decade at Warsaw University of Technology, new research began, led by Professor Andrzej Kraszewski. It regarded dispersed JPOAT system of information on the environment, later transformed into EcoInfoNet project [14].

This was a new concept of a dedicated analytical system that was created. It was going to allow both to make fundamental quantitative analysis and to implement dedicated solutions into the existing database system of information on the environment. The system had to comply with a number of assumptions, which are going to be described later in this article, but its fundamental task was to enable to conduct analysis, often unconventional, both in operating regime and in research tasks, which is a difficult objective, regarding such a vast amount of data.

As an introduction, it must be mentioned that the objective of this kind of research is not only collecting data, although it appears an important and necessary task. The phenomenon of air pollution is connected with continuous development of a society. Urban areas are inhabited by vast number of people and as a consequence industry and transport are highly developed. It makes them territories of the most significant degradation of natural environment. Pollution emitted this way imposes hazards on people's health, which is a kind of feedback. The costs of development, borne by citizens of these urban-industrial areas, should be appropriately minimalized by suitable management of environmental protection in a given region, taking into account the protection of social environment, as well as natural environment.

Issues under investigation, represent a new aspect of the use of statistical methods and models, often their development. The concept of data quality is closely related with modeling and forecasting – it is an interdisciplinary problem.

The development of numerical weather prediction, which has been most fully represented by E. Kalnay [15], [16] – one of the leading researchers in numerical forecasting – has become transport modeling development natural part. Statistics methods and models are integral parts of this development. The beginnings of numerical weather prediction are associated with the names of Bjerknes, who in 1904 defined the concept of forecasting, and Richardson (1922). The pioneers of modern numerical forecasts are considered Chaney, von Neumann, Fjørtoft and Eliassen, who in 1950 built the first one-day weather forecast, using a numerical model.

The development of pollutant transport modeling proceeded differently. Szepesi (1989) distinguishes three phases of development. The years 1920 to 1947, the first phase, were associated primarily with theoretical foundations. In the experimental phase (years 1948–1968) a particular attention was paid to the turbulent motion and in-depth understanding of the rules governing the spread of pollutants. The modern phase, the beginning of which dates back to 1968, is characterized by the development of numerical solutions.

In the last two decades of the twentieth century in the modeling of pollutant transport appeared trends towards ever fuller use of available meteorological data and experience of the discipline. The result is a significant improvement in both simple models (e.g. CALMET, AERMOD) and the development of a comprehensive model, taking into account the description of the meteorological and complex chemical and photochemical transformations chain (e.g. MM5-EURAD, MC2-AQ, UAM-V, WMF). These models are currently used in a wide range of scales, from global (e.g. Canadian GEM-AQ) on the scale of agglomeration – mezoscale.

Statistical weather, as a specific direction, owe their owes its development mostly to MOS model. The origins of MOS models are associated with the names of Howcroft (1971) and Glahn and Lowry (1972), who created the base model, "forgotten" for further

twenty years. The basis for the construction models are linear regression equations, which included a large number of independent factors (both meteorological and geographical parameters). The essence of the forecasting models of this kind is strong dependence of their results on the length of the historical data at the input. MOS models are used on a large scale.

At present, intensively extended, among others by Kalnay and Toth (1993, 1997, 2004), is ensemble forecasting (bundle, cluster forecasts or dynamic – statistical forecast), in which the Kalman filter has been applied, among others. Forecasts of this kind, from December 7, 1992, have been used by the National Meteorological Center. The essence of them is to use not a single prediction, but group of projections (bundles).

The origins of the use of statistical methods as self-direction in predicting the concentrations of pollutants in the atmosphere go back to the 1970s and are associated mainly with stochastic processes. McCollister and Wilson (1974) attempted to use the ARIMA model for forecasting daily episodes [19] of carbon dioxide and ozone concentrations in Los Angeles. Finzi (1979) attempted to predict sulfur dioxide ARIMAX model using as exogenous variables [12], wind speed and wind direction. Simpson and Layton (1983), using the Box Jenkins methodology, developed a prognostic model [25] of afternoon concentration increases in Brisbane. Inoue (1986) built a regression model estimating nitric oxide concentrations. The important date is 1995. Ziomas proposed a nitrogen dioxide concentrations model using meteorological data in Athens, while Ji Ping Shi and Harrison (1997) presented predictive model of nitrogen dioxide and oxides concentrations based on AR processes.

Other researchers, among whom should be mentioned Hernández (1992, 1999) and Chen (1998), proposed models based on stochastic processes used in air pollution forecasting, however, as so far, no comprehensive study of this issue has been published [24], [22]. Researchers have focused attention in a natural way around the metropolitan urban – industrial applications such as Paris, London and Madrid or the big metropolitan urban – industrial United States and Canada. In light of this trend it is worth noting the team focused in the center of Vilnius, Lithuania, in the years 1999–2005 was active in the field of science (Žičkus 1999–2003) expanding metropolitan area of research for the Nordic countries.

Modeling with the use of stochastic and exploratory techniques seems to be a perfect complement to balance models in operation activity of automatic networks monitoring conditions of the atmosphere. There are at least two areas where this research proves to be useful: automatic concentration monitoring combined with forecasts warning against excessive increase in pollution level and current analysis of data quality coming from monitoring network. Both these aspects, supported by a complex mathematical system, might become a mutually complementary, new area of use for quantitative methods, familiar and widely used in other areas, also dedicated to environmental engineering.

## ANALYTICAL SYSTEM ECO DATA MINER – EDM

***System structure overview***

The system is being created at Gdynia Maritime University, in cooperation with Warsaw University of Technology, Wroclaw University of Technology, Institute for Social Research, and other research centres. The project is a complex concept of creating

a system of warning forecasts against high concentrations of pollutants, with the use of stochastic methodology. It is an informatics implementation into an existing system, which monitors conditions of the atmosphere in urban and industrial agglomerations. It is a key system for air quality management and ultimately for pollution emission control.

The key element of EDM system structure is a subsystem of warning forecasts for high concentrations of pollutants, prepared by means of stochastic and exploratory methods. The process of constructing the warning forecast model was aimed to be created step-by-step, with the key stages being as follows:

1. Analysis of measurement data quality,
2. Creating models in the domain of time,
3. Identification of the model of warning forecast for excessive pollution levels.

The analysis of **measurement data quality** is carried out in accordance with own methodology proposal [4], [5], using robust estimators and being a complex solution for this issue. Attention to data quality measurement is an important issue, if not a key one, in a system of continuous atmosphere monitoring. Therefore this element was emphasized, proposing numerical implementation of both methods used in the European Union, the United States [11] as well as own methodology.

**Pollution forecast model** (short-term, multivariate) is based on principal components analysis by Hotteling (PCA). Based on the previous research, the model allowed to forecast the atmospheric conditions leading to high concentrations of pollutant from two to ten hours in advance [3], [7]. The idea of a system based on the combination of meteorological conditions enables to extend it further in a simple, factual way, turning it into a flexible tool for further research and for practical use in the scale of urban-industrial agglomerations [20].

**Warning system** is the key element of emission control in urban-industrial agglomerations. Combined with large-scale forecasts of meteorological conditions, it is going to allow for decreasing real dangers to people's health, resulting from pollution. Two classes of univariate models in time domain have been used: ARIMA and spectral, out of which, as proved by former studies, spectral models are more commonly used.

Another area to use the models in the domain of time in atmosphere pollution analysis is interpolation of missing data in measurement results [11], [7]. An argument in favour of this statement might be little requirements in terms of numerical implementation (the simplicity of implementation and short time needed for calculations). An assumption for this kind of models, and at the same time for forecasts received from them, is the inertia of the phenomenon, i.e. similarity of interpolated periods to previous periods, taking into account the character of the phenomenon (fundamental features, including its spatial presentation). Therefore, this approach becomes more advanced compared with simple interpolation functions.

The complexity of the proposed project can be noticed in creating an information system, whose task is to conduct the forecasting process, starting with the phase of measuring data quality assessment, through statistical methods implementation (which support analysis in operating mode), and local and mesoscale data integration into half-automatic forecasting system.

Flexibility of the system allows for its further extensions through dedicated models, which will realize new, currently not included, tasks in the future.

***System engineering***

The system is a dedicated analytical platform, which 'natively' uses local data received from atmospheric conditions monitoring networks and allows for integrating the data with the data from mesoscale systems. It operates in urban-industrial areas using research dedicated to environmental protection for air quality management systems.

There were implemented two main dedicated modules based on analytical platform: a module for measurement data quality assessment and warning forecasts, whose implementation was subsequent to the following studies:

- A study of methodology for measurement data quality assessment in automatic networks monitoring atmospheric conditions in an agglomeration as well as extension of previously used proposed methods [11], [6]. It is aimed at supporting the process of verification of historical data quality in time, spatial and frequency perspective;
- Creating a system warning against high concentration of pollutants, which allows to use local data fully and to integrate them with data from mesoscale models [3], [7].

Within detailed aims, supporting users must be mentioned in terms of ensuring the following:

(i) appropriate accuracy and precision of the research results,
(ii) appropriate time covering and completeness of measurement series,
(iii) comparability and reproducibility of tests in time and space .

***Analytical platform***

The design of the system was based on the assumptions that may be summarized by the following:

1) Flexibility – is realized by integrated IM modules and external EM modules – Figure 2,
2) Stability – realized by assuming that only MS Windows operating system mechanisms will be used while programming,
3) Scalability – external EM modules, DM dedicated modules – Figure 2,
4) Convenience for users – both advanced and those who need only a reporting tool- realized by user's interface,
5) Easy serviceability – the application is programmed in the way that allows user's problems to be identified in the widest and fastest way.

The analytical platform is the main element of the system. Its task is to integrate low-level modules (libraries and mathematical, statistical econometric and procedures) with dedicated modules (e.g. measurement data quality analysis). Multilayer structure of the application (Fig. 1) enables relatively simple expansion both on the side of dedicated external modules, which can be created in any language as classes, using EDM system in the function of name space, which contain cohesive environment of object system of database management and which make analytical methods available, as well as on the side of internal modules such as e.g. a subsystem of conducting surveys in the future.

Integrated modules: **Spreadsheet, Data, Reports, Results**, which can share both data streams and calculation results, are, from the user's point of view, cores of the application.

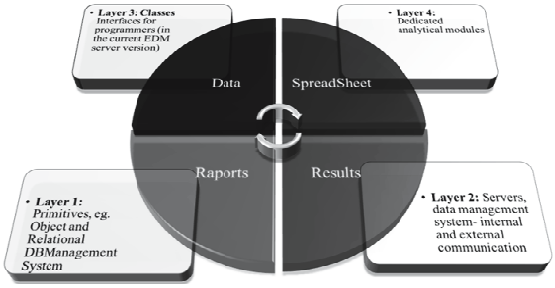


Fig. 1. Layers of EDM system application
Layer 1: Primitives, e.g. Object and Relational Database Management System
Layer 2: Servers, data management system- internal and external communication
Layer 3: Interface of classes for programmers (in the current EDM server version)
Layer 4: Dedicated analytical modules

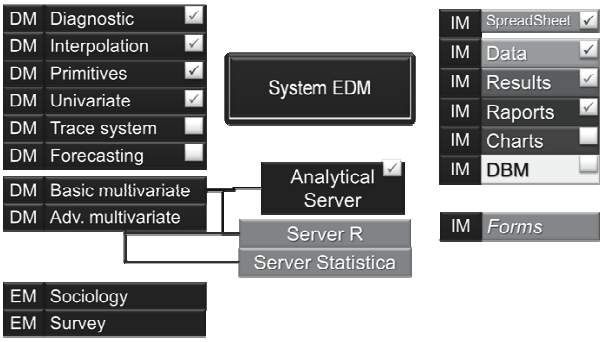


Fig. 2. EDM system internal structure

In terms of system operation, the core is an analytical server which in fact manages the whole system and shares its own interface (interface of server class) to be used by programmers to build external modules. The features which the system is subordinated to were taken into account in order to depict the structure of the server precisely from a user's point of view and in terms of objectives. It allows to show the capabilities of the analytical platform synthetically.

***Methods of outliers identification – characteristics***

Previously, data correctness was analysed mainly on the basis of researcher's experience. The person responsible for analysing measurement data, through their own experience and information about technical condition of equipment (measuring and transmission), manually rejected the incorrect data. This method is effective only for a small number of measuring data and a knowledgeable expert. In the case of continuous monitoring in many measuring stations, it turns out to be highly ineffective and there are many possibilities of making a mistake or overlooking. The increase in the amount of information makes it necessary to use quantitative methods.

The lack of both general and detailed suggestions about this issue leads to use simple methods in practical aspects, often in an incorrect way, without setting elementary foundations of their applications. An example can be a common application of the Three Sigma Rule, based on which it is assumed that outliers are those values which differ from the average ones by more than three standard deviations, *in plus* and *in minus*. Normality of distribution is not usually checked, either, along with chance of data variation, all of them leading to mistakes.

Daley [8] emphasizes the usefulness of statistical control cards in measurement data quality analysis [11] as well as in Bayesian analysis [11]. According to the EPA guidelines, it is suggested to use four statistical tests (extreme values – Dixon's Test for $n \leq 25$, Discordance Test for $n \leq 50$, Rosner's Test for $n \geq 25$ , Walsh's Test for $n \geq 50$) which are used to detect outliers, and additionally to compare current periods of time with corresponding periods of time from the past (e.g. concentrations in months of a given year compared to the same months of the previous year).

The second element of measurement data quality assessment is the case of missing data, which can have different reasons. The guidelines of General Directive set in detail a number of observation series which allows determining the quality of a given series. This kind of approach, useful in practical aspects, allows for carrying out analysis only in frequency-based approach. In statistical forecast, data with some missing values are useless in many cases, and using incomplete data might lead to incorrect results.

As a consequence, it leads to uncertainty about the likelihood of data, and thereby about analysis results quality, including forecasts, as well as accuracy of conclusions based on them. Methodology proposed later in this study seems, according to empirical results, to complement sufficiently the shortages which appear during determination and analysis of measurement data quality.

***Classic methods with implementation expansion in EDM***

*Grubbs's test* is used to detect observations with gross error outliers. Its advantage is applicability to samples of small and big size. The assumption of the test is normality of the variable distribution, which can be verified in EDM system by Shapiro-Wilk-Francia test, with Royston algorithm, and for big samples, over 5000 observations, with methods that are based on Kolmogorov – Smirnov – Lilliefors tests and Jarque-Bera test.

Classical procedure of Grubbs' test requires formulating the H0 hypothesis: that there are no outliers to the alternative H1 hypothesis: that there is at least one gross error outlier in the data set.

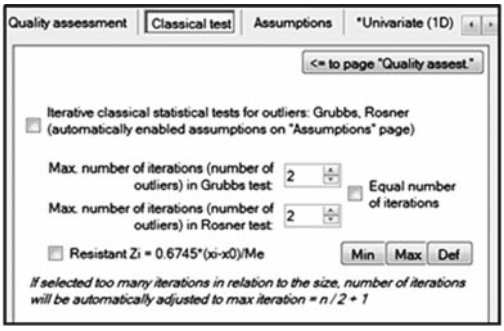


Fig. 3. Classic outliers tests implementation in EDM system

Test statistic is defined as:

$$G = \frac{\max_{i=1,\ldots,N}|x_i - \bar{x}|}{S} \tag{1}$$

with $\bar{x}$, S and N denoting the sample mean, standard deviation and sample size, respectively. Critical values, with the assumption of two-sided area, are determined as follows:

$$G_\alpha \geq \frac{N-1}{\sqrt{N}} \sqrt{\frac{t^2_{\alpha/2N,N-2}}{N-2+t^2_{\alpha/2N,N-2}}} \tag{2}$$

P level is approximated with t-distribution:

$$p = 2 * t(T, v = N - k), \text{ where } T = \sqrt{\frac{N(N-2)G^2}{(N-1)^2 - NG^2}} \tag{3} [14]$$

While implementing classical tests for outliers in EDM system, the algorithm contains procedures that allow identifying a given number of measurements, suspected to be untypical. In Grubbs's test, in each of the subsequent iteration, empirical distribution is reduced by the number of previously identified observations – one in each step. An advantage of the solution for big samples, that might consist of up to tens of thousands observations, is high efficiency. On the other hand, a disadvantage, mentioned in some studies, is a possibility of categorizing some correct data as untypical in the case of big number of iterations, which must be taken into account while carrying out the test.

*Rosner's test* is one of the tests recommended by EPA [11], [28], [11]. Originally the procedure was applied for small samples only – starting with the size of 10 and 25 outliers after 500 observations. Rosner's studies [31], [15] and Rong Chun Yu's ones [28] allowed to determine critical values, in practical aspects, for any size of the sample, owing to which the procedure has been widely applied.

The test procedure is quite complicated. A detailed description can be found in specialist literature [11], [31], [11]. Generally, a test statistics is used to detect potential outliers, according to the formula:

$$R_i = \frac{|x^{(i-1)} - \bar{x}^{(i-1)}|}{S^{(i-1)}} \tag{4}$$

with: *i* denoting a subsequent iteration which is the initially declared number of outliers in a distribution. An input parameter is a given number of outliers, on the contrary to the classical Grubbs' test which checks single observations in one course. Classical approach to Rosner's test requires using tables to determine critical values and it limits the number of observations to 500. In EDM system λ values approximation was applied [31], [15] where:

$$\lambda_i = \frac{(n-i)t_{n-i-1,p}}{\sqrt{\left(n-i-1+t^2_{n-i-1,p}\right)(n-i+1)}} \tag{5}$$

with:

$$p = 1 - \left[\frac{2\alpha}{2(n-i+1)}\right] \tag{6}$$

$\alpha$ is significance level of one-sided hypothesis.

***Proposed methodology***

The aim of the proposed methods is assistance while detecting potential outliers. The final decision of eliminating observation results from data set is taken by a human. *The idea* of the method was based on analysis of untypical observations (influential observations, distance observations or outliers).

The method suggested in this study allows to detect the presence of a situation which does not comply with the nature of a given phenomenon and to assess how the removal of outliers might affect the whole distribution of pollutants in univariate approach.

The key issue of the solution is the estimation of three main measures, which enable to classify the causes of identifying an observation as incorrect or untypical:

*DFITS measure* – indicates highly untypical nature of an observation, without listing its causes.

*Mahalanobis distance* – allows to estimate the distance between the result listed by DFITS and distribution of dependent variables (e.g. concentration). Therefore, it is possible to answer the question whether the cause of identification is connected to the dependent variable.

*Cook's distance* – allows to estimate the distance between a measurement and centroid, which is a point of reference in a multidimensional space of independent variables (e.g. meteorological measurements, data from balance models or predicators identifying periodicity).

In the course of research the stochastic-exploration apparatus identifying gross errors has been significantly expanded both on the side of measures and methodology in relation to the basic idea. Its full description goes beyond the scope of this description; therefore some key issues will be presented below.

It is assumed that outliers (in relation to univariate distribution and those not caused by changes in atmosphere) will allow to identify abnormalities in measurements not resulting from natural causes. To achieve its aim, the method requires realizing certain, listed below, stages:

*The first stage* is identification of probability density function of pollutants concentration values;

*The second stage* consists in determining theoretical and empirical percentiles of the examined distribution. Theoretical percentiles were assumed to be independent variable ($x_i$), empirical – to be dependent one ($y_i$), which allows to determine the residual in the form of:

$$e_i = y_i - \hat{y}_i \tag{7}$$

where: $y_i$ – empirical percentiles of linearized, identified distribution function, $\hat{y}_i$ – theoretical empirical percentiles of a linearized distribution function, in the form of $\hat{y}_i = a + bx_i$. Distribution functions linearized this way allow to detect observations, generally called outliers, based on projection matrix;

*The third stage* is a proper analysis, i.e. outliers identification;

*The last stage* is a final verification of the observations distinguished in the previous stage, made by a human.

The recommended method allows to detect outliers in univariate distribution. Figure 4 presents the model stages listed above, whose application can be modified according to aims.

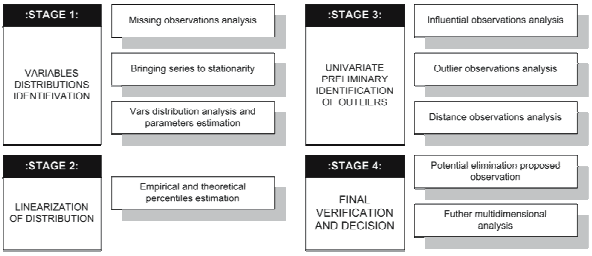


Fig. 4. Univariate data quality analysis stages in EDM system – base scenario

A series of univariate data quality analysis systems was implemented in EDM system (Fig. 4), in order to be able to carry out analyses in different approaches, including (i) based on real measurements, (ii) based on linearized univariate distributions, (iii) based on Fourier's models.

Each approach is aimed at different purpose and they can be chosen both by an operator in a manual mode, with possibility of setting various additional parameters, as well as in an automatic mode, generated for the needs of preparing reports.

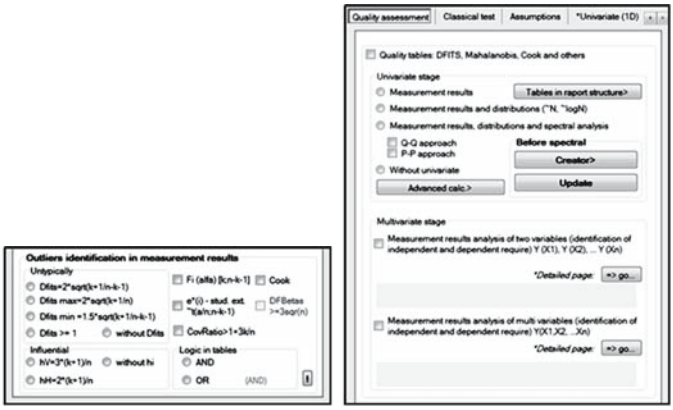


Fig. 5. Data quality analysis tools in toolbox

Multivariate approach is the next step, after univariate analysis, expanded by estimation of effects of different factors, e.g. atmospheric, on values that are initially marked as outliers.

Therefore, a character of variables is identified (endogenous, exogenous, or factors of different kind). For a series representing each variable, the quality will be assessed. In the second stage, a period of data is chosen to be analysed. This element is important, as multivariate approach assumes linear dependence between an independent and dependent variables.

Multivariate approach was implemented in two approaches (Figs 5 and 6): in pairs, where a particular dependent variable is analysed with each independent variable separately and in the approach of combined influence of all given independent variables on a particular dependent variable.

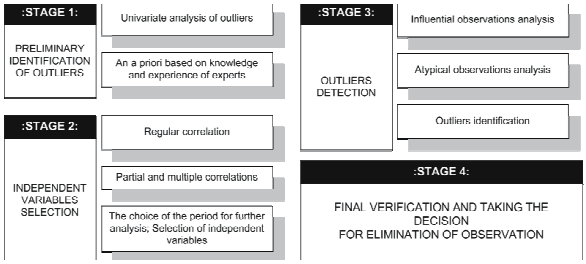


Fig. 6. Multivariate data quality analysis stages in EDM system – base scenario

***Methodological outline***

The main identification tool are standardized residuals in the form of

$$e_i' = \frac{e_i}{S} \tag{8}$$

with

$$S^2 = \frac{e^T e}{n-k-1} \tag{9}$$

being a classic parameter estimator $\sigma^2$.

Standardized residuals in the form of

$$e_i^* = \frac{e_i}{S\sqrt{1-h_i}} = \frac{e_i'}{\sqrt{1-h_i}} \tag{10}$$

or

$$e_i^* = \frac{e_i}{S_{(i)}\sqrt{1-h_i}} = \frac{e_i}{\sqrt{S_{(i)}^2 1-h_i}} = e_i\sqrt{\frac{S^2}{S_{(i)}^2}} = e_i\sqrt{\frac{(n-k-1)-1}{(n-k-1)-e_i^{*2}}} \tag{11}$$

with: $S_{(i)}$ denoting estimation of standard deviation of $\sigma$ random component, after eliminating $i^{\text{th}}$ observation. Standardized residuals are not stochastically independent. Residuals in the form of (10) and (11) are respectively called *internally* and *externally* studentized.

Internally studentized residuals have approximately t-Student distribution with *n-k-1* degrees of freedom, while externally studentized ones have exact t-Student distribution with *n-k-2* degrees of freedom. Outliers are characterized by externally studentized residuals.

The main tool used to detect *influential observations* is projection matrix:

$$H = X(X^T X)^{-1} X^T \tag{12}$$

Diagonal elements of the *h*i matrix are influential values. They have the key role in the analysis. They determine the influence of particular observations on regression model parameters assessment. A theoretical value of an explanatory variable can be calculated by using the equation:

$$\hat{y}_i = \sum_{j \neq i} h_{ij}\, y_j + h_i y_j \tag{13}$$

Projection matrix is symmetric and idempotent, i.e. the following dependences are fulfilled: $H = H^T$ and $H^2 = H$.

These properties lead to the following inequality: $h_i = h_i^2 + \sum_{j \neq i} h_{ij}^2 \geq h_i^2$ which implicates the second important property of the *H* projection matrix:

$$0 \leq h_i \leq 1 \tag{14}$$

The effect of $i^{th}$ observation on the change of theoretical values of the explained variable depends only on the value of the residual and on $i^{th}$ influential value. High $h_i$ values, close to unity, influence theoretical values of this variable significantly, even if $i^{th}$ residual is small. Eliminating $i^{th}$ observation, for which influential value is high, might change regression analysis results remarkably. Therefore, $h_i$ might be considered a good influential observations indicator.

Property (14) allows to make simple interpretation of influential values, which is an advantage in terms of analysis in the case when a reference model, e.g. of regression, considers a free term $h_i \geq \frac{1}{n}$.

Influential values can be treated as meters of distance of certain observations from reference vector, e.g. mean values vector $\{\bar{X}_1, \bar{X}_{2,\ldots,} \bar{X}_k\}$.

An indicator which measures both untypicality and influentiality of an observation – standardized difference between theoretical values, is an indicator defined as

$$DFITS_i = e^*_{(i)} \sqrt{\frac{h_i}{1-h_i}} \tag{15}$$

It is a standardized value of an effect made by eliminating $i^{th}$ observation. Another measure of total assessment, whose structure is based on the matrix $S^2(X^TX)^{-1}$ with, $S^2_{(i)}(X^TX)^{-1}$ is COVRATIO measure, according to

$$COVRATIO = \frac{\det\left\{S^2_{(i)}\left(X^T_{(i)}X_{(i)}\right)^{-1}\right\}}{\det\left\{S^2_{(i)}(X^TX)^{-1}\right\}} = \frac{1}{\left[\frac{n-k-1}{n-k}+\frac{e^{*2}_{(i)}}{n-k}\right]^k (1-h_{ij})} \tag{16}$$

This coefficient is not normalized [10]. Value "1" is used for interpretation. If the result is smaller than unit, it implies the presence of untypical value and possible improvement of estimator efficiency.

The next element of regression diagnostics are two measures of distance: Mahalanobis distance and widely applied *Cook's distance*, represented by the equation

$$d_i = \frac{(b_{(i)}-b)^T X^T X(b_{(i)}-b)}{(k+1)S^2} \tag{17}$$

This distance takes into account both the explained variable and explanatory ones. Its value is a standardized distance of regression parameters vector, after eliminating $i^{th}$ observation $b_{(i)}$ from regression parameters vector $b$ for the whole set of data. It measures the effect of single observations on the change of all regression parameters. It enables ordering observations in terms of power of influence on parameters assessment.

An observation is classified as highly influential if the Cook's distance related to it is high. Function $d$ has ellipsoidal shape in a space with $k+1$ dimensions and the centre in $b$ point, so parameters vector for regression determined for the full set of observations in the centre of these ellipsoids.

## EXAMPLES OF IMPLEMENTATION

*Data* – calculations were made for results of one-hour measurements from an automatic, traffic-related air pollutants monitoring station (marked as WaK). in Warsaw in 2009 – the variables represent series of one-hour measurement results – time series. Initial calculations allowed to choose contrasting variables, both having relatively low and high information range (Vs), with the criterion of the highest possible observation number (Ni), both in warm and cold season – Table 1.

Two variables were chosen for further analysis: $NO_2$ and toluene, in two months: July and November. An additional criterion to choose the variables was monthly course of correlation coefficients between traffic volume and variables. Correlation levels in July and November can be classified as 'typical', for $NO_2$ at the level of 0.54 and 0.57 and for benzene 0.32 and 0.30 respectively – Figure 7.

Missing data were interpolated based on previous studies [6], owing to which it was possible to carry out a full harmonic analysis for three out of four variables – Table 2.

It was impossible to interpolate measurements of benzene in November because of missing data distribution, which made it impossible to use previously analysed methods.

Table 1. WaK station variability estimation

| Variable | NBD | Nbd % | x0 | Sn | x0_W | x0_U | X0_W-X0/X0 % | X0_U-X0/X0 % |
|---|---|---|---|---|---|---|---|---|
| $NO_2$ [µg/m$^3$] July | 744 | 100.0 | 52.98 | 25.38 | 52.18 | 51.39 | -1.5 | -3.0 |
| Benzene [µg/m$^3$] July | 744 | 100.0 | 1.53 | 0.96 | 1.46 | 1.42 | -4.6 | -7.2 |
| $NO_2$ [µg/m$^3$] November | 720 | 100.0 | 49.44 | 29.12 | 47.69 | 45.9 | -3.5 | -7.2 |
| *Benzene [µg/m$^3$] November* | *702* | *97.5* | *4.88* | *2.49* | *4.81* | *4.71* | *-1.4* | *-3.5* |

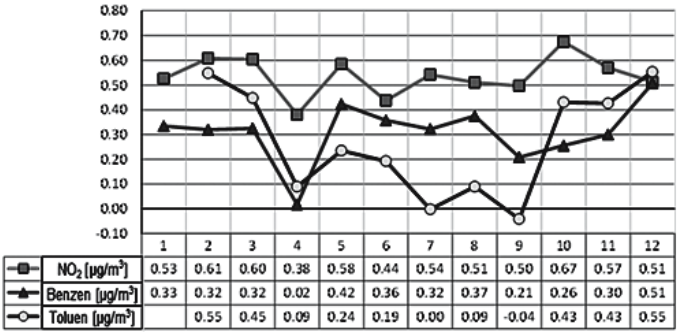


Fig. 7. 2009 monthly Pearson's correlation coefficient between traffic volume and pollutants

Table 2. Variables moments of distributions

| Variable | NBD | Nbd % | x0 | Sn | x0_W | x0_U | X0_W-X0/X0 % | X0_U-X0/X0 % |
|---|---|---|---|---|---|---|---|---|
| $NO_2$[µg/m$^3$] July | 744 | 100.0 | 52.98 | 25.38 | 52.18 | 51.39 | -1.5 | -3.0 |
| Benzene[µg/m$^3$] July | 744 | 100.0 | 1.53 | 0.96 | 1.46 | 1.42 | -4.6 | -7.2 |
| $NO_2$[µg/m$^3$] November | 720 | 100.0 | 49.44 | 29.12 | 47.69 | 45.9 | -3.5 | -7.2 |
| *Benzene[µg/m$^3$] November* | *702* | *97.5* | *4.88* | *2.49* | *4.81* | *4.71* | *-1.4* | *-3.5* |

***The model***

As an example of measurement data quality assessment univariate analysis was chosen, which was based on spectral (harmonic) model, i.e. the model is a reference point of the analysis, so that the detailed analysis is made on its basis. One of the advantages of this kind of approach is that identification of different kinds of periodicity sources, e.g. daily cycle or monthly cycle, is made in the model, maximum 24 harmonics in one course and any other time-consuming calculations are not necessary to identify significant harmonics.

Each periodic function with $n$ period can be described as a sum of harmonics, sinusoidal and cosine functions with periods $n/i$ ($i = 1, 2, ..., n/2$), that is $\theta_i = \left(\frac{2\pi}{n} i\right)$ which allows to describe the model form as:

$$x_t = f(t) + \sum_{i=1}^{\frac{n}{2}} \left[\alpha_i sin\left(\frac{2\pi}{n} it\right) + \beta_i cos\left(\frac{2\pi}{n} it\right)\right] \quad (18)$$

with: $i$ – denoting harmonics number; f(t) – reference level (function, e.g. lineal trend or mean level) $\alpha_0$, $\alpha_i$, $\beta_i$ – estimated parameters:

$$a_0 = \frac{1}{n}\sum_{t=1}^{n} x_t;\ a_i = \frac{2}{n}\sum_{t=1}^{n} x_t\ sin\left(\frac{2\pi}{n} it\right), i = 1, \ldots, \frac{n}{2} - 1;\ b_i = \frac{2}{n}\sum_{t=1}^{n} x_t\ cos\left(\frac{2\pi}{n} it\right) \quad (19)$$

with: $x_t$ – denoting a series of results, $a_0$, $a_i$, $b_i$ – assessment of $\alpha_0$, $\alpha_i$, $\beta_i$ parameters.

The model identified in the above way is a reference point for quality assessment, with use of the measurements and methods described above.

## RESULTS OF THE RESEARCH, INTERPRETATIONS AND CONCLUSIONS

In July 2009 one episode of $NO_2$ concentration value was noticed (172,5 µg/m$^3$; 2009-07-30 10:00 pm; obs. 719,8) as well as the situation, when the concentration decreased, in relatively short time of about 4 hours, from about 21 µg/m$^3$ to 9,3 µg/m$^3$ (2009-07-17 04:00 am; obs. 393, Fig. 8) to increase within the next four hours to over 20 µg/m$^3$. The good representation of the AH harmonic model in Figure 9 is worth noticing. It enables precise analysis.

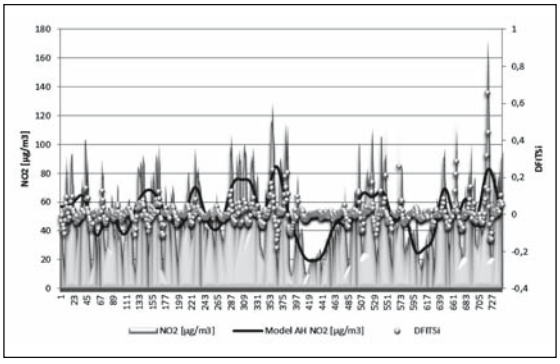


Fig. 8. Data quality analysis of NO2 with QAAH1 method for July 2009

Both situations were correctly identified by data quality assessment system as potentially outlying, both with classic methods (Table 4) and the suggested QAAH1 method (Table 3). Higher precision of the supplied information must be noticed for the suggested method. Classic approach (Rosner's and Grubbs's tests) informs only that single

observations are suspected to be outlying. QAAH1 method gives additional information (Table 3) about their character. Fi statistics completes the information with statistical significance of the observations 'around' the identified one. High values of Cook's distance prove a cause 'beyond' the scope of univariate approach, which is confirmed by Mhn values. High values of concentrations should not be interpreted as outliers, but only as distant ones. The conclusions are proved by 'not significant' influential values of hi.

Table 3. Detailed measures of $NO_2$ data quality in QAAH1 method (July 2009)

| $NO_2$ | $t$ | $h_i$ | $e^*_{(i)}$ | $DFITS_i$ | *COVRATIO* | *Z DFITS*$_i$ | *Cook* $d_i$ | ***Mahalanobis*** | $F_i$ |
|---|---|---|---|---|---|---|---|---|---|
| 102.8 | 716 | 0.0065 | 0.9771 | 0.0792 | **1.0066** | 0.0352 | 0.0031 | 3.8542 | **3.8744** |
| 116.4 | 717 | 0.0097 | 1.4564 | 0.1445 | **1.0098** | 0.0541 | **0.0104** | 6.2456 | **6.2986** |
| 140.6 | 718 | 0.0174 | 2.3785 | 0.3164 | **1.0177** | 0.0913 | **0.0497** | 11.9210 | **12.1157** |
| **172.5** | 719 | 0.0312 | 3.6539 | **0.6557** | **1.0322** | 0.1443 | **0.2114** | 22.1809 | **22.8644** |
| 154.6 | 720 | 0.0229 | 2.8814 | 0.4414 | **1.0235** | 0.1121 | **0.0965** | 16.0347 | **16.3888** |
| 116.1 | 721 | 0.0097 | 1.3094 | 0.1294 | **1.0098** | 0.0486 | **0.0084** | 6.1866 | **6.2386** |

Table 4. Classic test of data quality assessment: modified iteration tests of Grubbs and Rosner for $NO_2$ concentrations (July 2009)

| $NO_2$ | Iteration | **G#** | p(G)# | X0# | $S_n$# | $\mathbf{R_i}$ | Lambda(C) | p($R_i$) |
|---|---|---|---|---|---|---|---|---|
| 130.1 | 4 | 3.1580 | 1.0000 | 52.5601 | 24.5532 | 3.1580 | 3.9650 | 0.6802 |
| 140.6 | 3 | 3.5526 | 0.5365 | 52.6788 | 24.7485 | 3.5526 | 3.9654 | 0.2355 |
| 154.6 | 2 | 4.0693 | 0.0639 | 52.8160 | 25.0129 | **4.0693** | **3.9657** | **0.0314** |
| **172.5** | 1 | **4.7097** | **0.0031** | 52.9768 | 25.3783 | **4.7097** | **3.9661** | **0.0016** |

Apart from information referring to a particular 'incident', concentrations diagrams versus $DFITS_i$ (Fig. 9) and $F_i$ (Fig. 10) are especially valuable. The former one depicts clearly identified potential uncertain measurements as a 'continuous' chain, similarly to the nearest neighbours method in concentration analysis. However, the chain is not 'torn' and there is no measurement result whose *DFITS* value would differ remarkably (more than 1), which means that there are no outliers. This is confirmed by graph $F_i$ in which the same kind of 'chain' is depicted.

$NO_2$ measurement results from November 2009 are different to those from July 2009. No potentially hazardous phenomenon was identified. This situation is depicted by Figure 10. Low values of *DFITS*, under 0.4, prove the absence of dangerous deviations in univariate approach, considering the influence of 24 factors connected to changes in time (24 first harmonics).

Two-dimensional projection of DFITS versus Mahalanobis distance is an interesting approach (Fig. 11). It enables quick and precise assessment if there are any outliers present in the data set and if they are caused by deviations in distribution. Mhn is not normalized, but there were no values differing in order of magnitude from the stem.

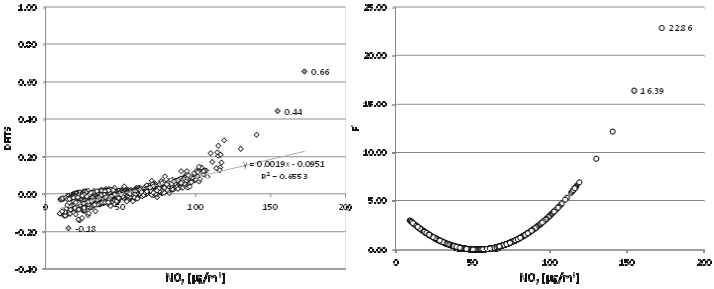


Fig. 9. NO2 versus DFITSi and Fi measurements with QAAH1 method for July 2009

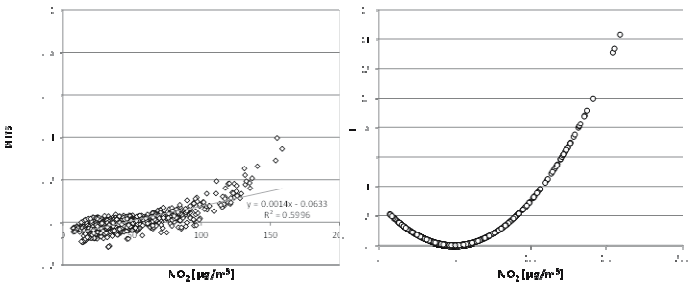


Fig. 10. NO2 versus DFITSi and Fi measures with QAAH1 method on November 2009

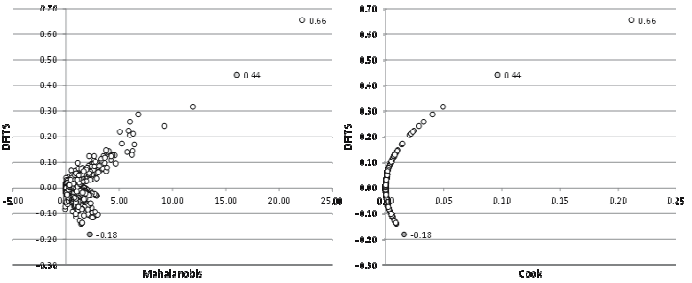


Fig. 11. Quality analysis measures of QAAH1 method: DFITSi v. Mahalanobis and Cook's distance 2d projection; NO2 July 2009

## SELECTED RESULTS – COMPARING TWO PERIODS IN GDANSK AGGLOMERATION

In the second example the data from 2010 are used, stations AM2 and AM5. Hourly series for measuring concentrations of NO, NOx, NO2 and PM10. Initial manual analysis was performed by the monitoring network operators, which identified the possibility of irregularity in data. The analysis was performed for the months of January – February (Fig. 12).

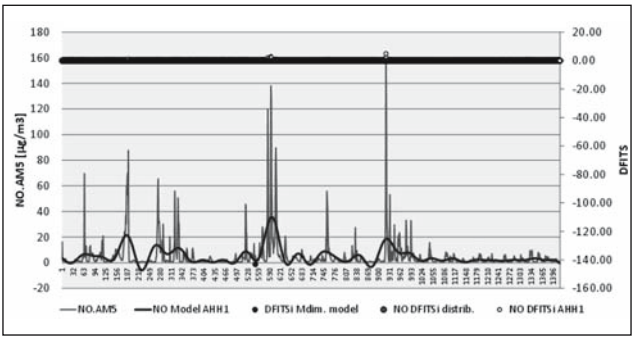


Fig. 12. DFITS values (multidimensional model, one-dimensional NO distribution analysis and based on the model AHH1) NO at the AM5 station in January / February 2010

In addition to model AHH1 DFITS values were calculated for the multidimensional model NO concentration (Y: NO.AM5 / [X]: SO2.AM5; NO2.AM5; NOX.AM5; CO.AM5; TEMP.AM5; PM10.AM5; NO2.AM2; PM10.AM2). Multivariate analyses indicate the point number 550, and confirm the initial findings.

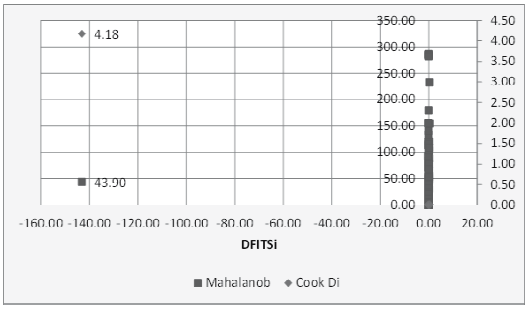


Fig. 13. DFITS values and the NO Cook's distance at the station AM5 for multidimensional model in the January / February 2010, together with the measurement of the number 550

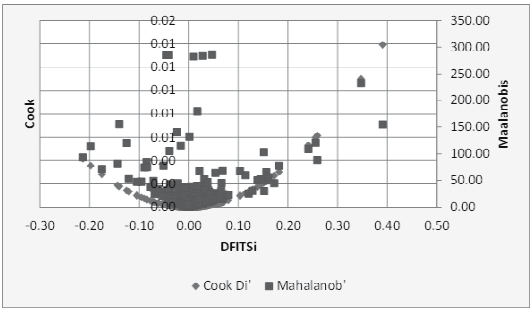


Fig. 14. DFITS values and the NO Cook's distance at the station AM5 for multidimensional model in the January / February 2010, together without the measurement of the number 550

Two figures above (Figs 13 and 14) indicate that the measurement of 550 is highly untypical because of the independent variables of the model (high values of Cook's distance) and the distribution (high values of Mahalanobis distance), but not in the sense of the spectrum (low values DFITS AHH1, Fig. 12). The results of the analysis indicate the need to investigate the causes of high value of DFITS in the model and on a multi--dimensional and one-dimensional distribution of both low values in the AHH1 DFITS analysis.

The measurements of independent variables, including PM10, may be the probable cause of identification. PM10 measurements are characterized by high variability.

Table 5. Quality diagnosis results for model: Y: NO.AM5/ [X]: SO2.AM5; NO2.AM5; NOX.AM5; CO.AM5; TEMP.AM5; PM10.AM5; NO2.AM2; PM10.AM2; January/February 2010

| Number | Value | UA_ Data | UB_ Distr. | UC_ QAAH1 | MA_ Pairs | MB_ All | IdOR |
|---|---|---|---|---|---|---|---|
| 594 | 137.999 | | 4.8% | 16.1% | 13.5% | | 25.9% |
| 921 | 162.21 | | 4.8% | 16.1% | 13.5% | | 25.9% |
| 585 | 120.029 | | | 16.1% | 13.5% | | 21.7% |
| 595 | 128.269 | | | 16.1% | 13.5% | | 21.7% |
| 922 | 113.278 | | | 16.1% | 13.5% | | 21.7% |
| 189 | 87.7725 | | | 16.1% | | | 14.1% |
| 608 | 89.6861 | | | 16.1% | | | 14.1% |
| **550** | **5.5678** | | | | | **19.4%** | **11.8%** |
| 1 | 162.21 | | | | 13.5% | | 7.6% |
| 1415 | 420.187 | 1.6% | | | | | 1.4% |

Both models, the results of which are shown in the tables above by explaining the variation of NO.AM5 by a large number of significantly related factors, accurately describe the natural causes of the formation of NO concentrations.

IdOR factor for the measurement of 550 is below 12%, which points to natural causes of irregularities. Independent variables explained the variation in the concentrations of NO in the measurement point. High values of DFITS were caused by an atypical situation of the independent variables, mainly PM10. This was confirmed by a detailed analysis.

## CONCLUSIONS AND FURTHER PROSPECTS

The application of quantitative methods, including stochastic and exploratory ones, in environmental studies, both empirical and methodological, does not seem to be sufficient in practical aspects.

Eco Data Miner analytic platform presented in this study, according to the received results and previous studies, seems to be effective in its application range. Abilities of the system, achieved by expanding a specific concept of platform structure enable both a wide

spectrum of application in practical aspects, and further development and improvements, based on future requirements, not known yet. Directions of future development can be divided into two main groups.

The first group includes technical aspects. Full implementation of graphics requires further works, including interactive ones. Reporting module requires further improvement and particular numerical procedures need to be corrected. A new module of multivariate analyses is under construction.

The second category includes the issues of long-term development of the system. Two directions seem to be key for this issue: (i) definite development of interpolation module, both one-dimensional and spatial ones, by designing new methods, and (ii) creating a module of dedicated control cards working in operator's regime in pseudo-real time, i.e. with sampling under one hour, which is an expansion of the presented concept of data quality assessment, including QAAH1 method.

*ACKNOWLEDGEMENTS*

*Designing Eco Data Miner system, in the part of Piotr Czechowski's work, was co-financed in 2006–2009 as a part of the ordered research project KBN „POL-POSTDOC II" Nr PBZ/MEiN/01/2006/53.*

## ABBREVIATIONS

| | |
|---|---|
| EDM | – *Eco Data Miner* – quantitative analysis system |
| AH | – *Harmonic Analysis* – harmonic (spectrum) analysis |
| QAAH1 | – *Quality Analysis based on Harmonic Analysis type 1* |
| p level | – probability level of null hypothesis rejection |
| IM | – *Integrated Module* – integrated module of EDM system |
| EM | – *External Module* – external module of EDM system |
| DM | – *Dedicated Module* – dedicated module of EDM system |
| NBD | – number of values without missing values |
| Nbd% | – percentage of correct measurements without missing values |
| x0 | – arithmetic mean (used in tables); other cases: $\bar{x}$ |
| min(x), max(x) | – minimal and maximal value |
| Sn | – standard deviations of the sample (used in tables); other cases: $\hat{S}$ |
| x0_W | – Winsorized mean |
| x0_U | – trimmed mean |
| $NO_2$ | – nitrogen dioxide measurement results |
| Mhn | – Mahalanobis distance |

## REFERENCES

[1] Badyda, A., Dąbrowiecki, P., Lubiński, W., Czechowski, P.O., & Majewski G. (2013). Exposure to traffic--related air pollutants as a risk of Airway Obstruction. *Advances in Experimental Medicine and Biology*, 755, 35–45, DOI: 10.1007/978-94-007-4546-9_5.

[2] Badyda, A., Dąbrowiecki, P., Lubiński, W., Czechowski, P.O., Majewski, G., Chciałowski, A., & Kraszewski, A. (2013). Influence of Traffic-Related Air Pollutants on Lung Function: Warsaw Study. *Advances in Experimental Medicine and Biology*, 788, 229–235, DOI: 10.1007/978-94-007-6627-3_33.

[3] Czechowski, P.O., & Kraszewski, K.A. (2009). Określenie horyzontu prognoz ostrzegawczych zanieczyszczeń atmosfery metodami stochastycznymi na przykładzie dwutlenku azotu w skali lokalnej – koncepcja systemu Eco Data Miner. *Zeszyty Naukowe Uniwersytetu Gdańskiego*, 38, 229–246.
[4] Czechowski, P.O. (2009). Mechanizmy oceny jakości danych pomiarowych w koncepcji systemu analitycznego Eco Data Miner, Gdańsk 2009.
[5] Czechowski, P.O. (2001). Ocena jakości danych pochodzących z sieci monitorującej stan atmosfery na przykładzie Aglomeracji Gdańskiej, Gliwice 2001.
[6] Czechowski, P.O. (2004). Statystyczna interpolacja braków danych pomiarowych [In.] *Raport 2003 o stanie zanieczyszczeń powietrza atmosferycznego w Aglomeracji Gdańskiej*, Gdańsk 2004.
[7] Czechowski, P.O. (2004). System prognozowania wysokich stężeń dwutlenku azotu w atmosferze (na przykładzie Aglomeracji Gdańskiej), Warszawa 2004.
[8] Daley, R. (1991). Atmosferic Data Analysis, Cambridge University Press 1991.
[9] Domański, Cz., Pruska, K., & Wagner, W. (1998). Wnioskowanie przy nieklasycznych założeniach. Wydawnictwo Uniwersytetu Łódzkiego, Łódź 1998.
[10] Domański, C., & Pruska, K. (2000). Nieklasyczne metody statystyczne; PWE; Warszawa 2000.
[11] EPA, Data Quality Assessment: A Reviewer's Guide, EPA QA/G-9R 2000 & EPA/240/B-06/002. 2006.
[12] EPA.; Guidance for Data Quality Assessment; Practical Methods for Data Analysis; EPA QA/G-9 QA00 Update; EPA/600/R-96/084.2000, 2007.
[13] Finzi, G., Calori, G., & Tonezzer, C. (1993). A decision support system for air quality network design; DSS for air quality management. *Environmental Monitoring and Assessment*, 33, 101–114.
[14] Solak, M.K. (2009). Detection of multiple outliers in univariate data sets. *PharmaSUG 2009*: http://www.lexjansen.com/pharmasug/2009/sp/sp06.pdf.
[15] Iglewicz, B., & Hoaglin, D.C. (1993). How to Detect and Handle Outliers. ASQC Quality Press, Milwaukee 1993.
[16] Juda-Rezler, K., Reizer, M., & Oudinet, J.P. (2011). Determination and analysis of PM10 source apportionment during episodes of air pollution in Central Eastern European urban areas: The case of wintertime 2006. *Atmospheric Environment*, 45, 6557–6566.
[17] Kalnay, E. (2003). Historical overview of numerical weather prediction [In.] Atmospheric modeling, data Assimilation and Predictability, Cambridge Univeristy Press 2003.
[18] Kalnay, E. (2004). Postprocessing of Numerical Model Output to Obtain Station Weather Forecasts; Statistical Forecasting: with NWP. Working papers Meteo Course, 2004.
[19] Kraszewski, A. (2002). Ekoinfonet – Information System for State Monitoring of Environment in Poland. *EnviroInfo Vienna*, 1, 117–124, Viena 2002.
[20] Łobocki, L. (2003). Wskazówki metodyczne dotyczące modelowania matematycznego zarządzania jakością powietrza; Ministry of Environment; Chief Inspectorate of Environmental Protection; Warsaw 2003.
[21] Madany, A. (1997). Air quality simulation models in Poland. *Quarterly Journal of the Hungarian Meteorological Service*; 101, 1, 33–43.
[22] Majewski, G., Czechowski, P.O., Badyda, A.J., & Rogula-Kozłowska, W. (2013). The Estimation of Total Gaseous Mercury Concentration (TGM) Using Exploratory and Stochastic Methods. *Polish Journal of*
[8] *Environmental Studies*, 22, 3, 759–771.
[23] Majewski, G., Czechowski, P.O., Badyda, A., Kleniewska, M., & Brandyk, A. (2013). Ocena stężenia całkowitej rtęci gazowej (TGM) na terenie stacji tła regionalnego Granica-KPN (województwo mazowieckie, Polska) w latach 2010–2011. *Rocznik Ochrona Środowiska/Annual Set The Environment Protection*, 15, 1302–1317.
[24] Majewski G. (2009). Study of Particulate Matter Pollution in Warsaw Area. *Polish Journal of Environmental Studies*, 18, 2, 293–300.
[25] McCollister, G., & Wilson, K.; Linear stochastic models for forecasting daily maxima and hourly concentrations of air pollutants. *Atmospheric Environment*, 9, 417–423.
[26] Ostasiewicz W., [Ed.] (1998). Statystyczne metody analizy danych; Wydawnictwo Akademii Ekonomicznej im. Oskara Langego; Wroclaw 1998.
[27] Podawca, K., & Rutkowska, G. (2013). Analiza przestrzennego rozkładu typów zanieczyszczeń powietrza w układzie dzielnic m.st. Warszawy; *Rocznik Ochrona Środowiska/Annual Set The Environment Protection*, 15, 2090–2107.
[28] Rong Chun Yu, Hee Wen Teh,. Jaques, P.A., Sioutas, C., & Froines J.R.(2004). Quality control of semi-continuous mobility size-fractionated particle number concentration data. *Atmospheric Environment*, 38, 3341–3348.

[29] Siewor, J., Tumidajski, T., Foszcz, D., & Niedoba, T. (2011). Prognozowanie stężeń zanieczyszczeń powietrza w GOP-ie modelami statystycznymi. *Rocznik Ochrona Środowiska/Annual Set The Environment Protection*, 13, 1261–1274.
[30] Simpson, R., & Jakeman, A.(1984). A model for estimating the effects of fluctuations in long term meteorology on observed maximum acid gas levels. *Atmospheric Environment*, 18, 8, 1633–1640.
[31] Sioutas, C., & Rosner, B. (1983). Percentage points for a generalized ESD many-Outlier Procedure. *Technometrics* 25, 165–172.
[32] Tarkowski, R., Sroczyński, W., Luboń, K., Wdowin, M. (2012). Wstępne wyniki testu aparatury do ciągłego pomiaru stężenia CO2 w powietrzu glebowym na stanowisku Szczawnicy; *Rocznik Ochrona* [8] *Środowiska/Annual Set The Environment Protection*, 14, 930–944.
[33] Trapp, J.A. [Ed.]. (1996). Wstępne wyniki badań nad zmiennością średniej miesięcznej temperatury powietrza w Gdańsku w latach 1851–1995, Wydawnictwo DJ 1996.

# LIST OF REVIEWERS

*Archives of Environmental Protection* relies upon the commitment, expertise and judgment of its reviewers to maintain the high standard of research it publishes. Here we list the names of all those who have submitted one or more reviews for us between November 2012 and November 2013. We know that acting as a reviewer for *Archives of Environmental Protection* entails significant time and effort and we very much appreciate the support we have received from all these colleagues.

- Anielak Anna
- Antonkiewicz Jacek
- Apostoluk Wiesław
- Aranowski Robert
- Barabasz Wiesław
- Bodzek Michał
- Ciepał Ryszard
- Cwalina Beata
- Czamara Alicja
- Czaplicka Marianna
- Dunalska Julita
- Falandysz Jerzy
- Filipek Tadeusz
- Floryszek-Wieczorek Jolanta
- Gawdzik Andrzej
- Gierak Andrzej
- Gierycz Paweł
- Gonet Sławomir
- Góralczyk Stefan
- Grabas Kazimierz
- Gryta Marek
- Janosz-Rajczyk Marta
- Jarosiński Andrzej
- Jóźwiakowski Krzysztof
- Juda-Rezler Katarzyna
- Kabsch-Korbutowicz Małgorzata
- Kalembasa Dorota
- Kalembasa Stanisław
- Karczewska Anna
- Kleiber Tomasz
- Kliś Czesław
- Kołota Eugeniusz
- Kołtuniewicz Andrzej
- Konieczyński Jan
- Kordylewski Włodzimierz
- Kostecki Maciej

- Koszelnik Piotr
- Kowalik Piotr
- Kucharski Mariusz
- Kulig Andrzej
- Kyzioł-Komosińska Joanna
- Labus Krzysztof
- Lach Joanna
- Łączny Marian
- Łebkowska Maria
- Liwarska-Bizukojć Ewa
- Małachowska-Jutsz Anna
- Mazierski Jerzy
- Mazur Marian
- Miksch Korneliusz
- Mioduszewski Waldemar
- Mocek Andrzej
- Morawski Antonii
- Mrowiec Maciej
- Musialik-Piotrowska Anna
- Myszograj Sylwia
- Nadziakiewicz Jan
- Nawalany Marek
- Nowak Arkadiusz
- Obarska-Pempowiak Hanna
- Oleszek-Kudlak Sylwia
- Pacha Jerzy
- Pacyna Józef
- Pająk Tadeusz
- Palowski Bernard
- Pawełek Jan
- Pawlak-Kruczek Halina
- Pawlik-Skowrońska Barbara
- Piaścik Marek
- Piecuch Tadeusz
- Pietr Stanisław
- Piotrowska-Seget Zofia
- Płaza Grażyna
- Polechoński Ryszard
- Popek Włodzimierz
- Postrzednik Stefan
- Preisner Leszek
- Protasowicki Mikołaj
- Quant Bernard
- Rak Janusz
- Robak Małgorzata
- Rogula-Kozłowska Wioletta

- Rosik-Dulewska Czesława
- Rosiński Marian
- Rosińska Agata
- Rostański Adam
- Rzętała Mariusz
- Sądej Wiera
- Ściążko Marek
- Siepak Jerzy
- Siuta Jan
- Skiba Stefan
- Sozański Marek
- Spiak Zofia
- Staszewski Tomasz
- Steliga Teresa
- Strugała Andrzej
- Surmacz-Górska Joanna
- Suschka Jan
- Szalińska van Overdijk Ewa
- Szetela Ryszard
- Tabiś Bolesław
- Tomaszek Janusz
- Twardowska Irena
- Ulfig Krzysztof
- Weber Jerzy
- Wiatkowski Mirosław
- Wielgosiński Grzegorz
- Wiśniewski Ryszard
- Wolny Lidia
- Wysocka Małgorzata
- Wysokiński Lech
- Wyszkowska Jadwiga
- Zawadzki Jarosław
- Żeliński Jacek
- Zielewicz Ewa
- Żukowski Witold
- Zwodziak Anna
- Zygmunt Bogdan