EnviroInfo 2007 (Warschau)
Environmental Informatics and Systems Research
Copyright © Shaker Verlag, Aachen 2007. ISBN: 978-3-8322-6397-3

eGovernment Services in Environment - Automate Data Quality Assessment - Czech Republic Approach

Miloslav Hlaváček[1], Michal Hejč[2], Jiří Hřebíček[3]

Abstract

Imprecision of data is an important characteristic feature of environmental monitoring, particularly at all kinds of fixed and moving sensor networks. Imprecision of primary data is an important characteristic feature of environmental monitoring. When making evaluations, conclusions and the decisions from collected environmental data, we have to assess data quality not to make fatal mistake in environmental management. The important task of eGovernment services in the Czech Republic is to deal with the primary data uncertainty and using knowledge about data quality to assure public access to quality environmental information (Directive 2003/4/EC) and provide added value for the citizens and businesses (end users and decision makers). This will reduce the risk of inequality information in eGovernment services and wrong decision in environmental management.

At present, new approaches and methodologies of eGovernment services handling with environmental data uncertainty are explored. The paper presents comparison of such new approach developed in the Czech Republic against European Environment Agency and U.S. Environmental Protection Agency approaches. The new approach to information management with data quality assessment brings for the Visegrad countries more flexible way of dealing with all kinds of uncertainty in eGovernment services. There are presented better results and less workload then standard approaches in the paper.

1. Introduction

Imprecision and uncertainties of environmental data and information is an important characteristic feature of environmental monitoring and management. In the context of decision support there are two typologies of uncertainty (van Asselt, 1999). First, variability and lack of knowledge are broken down into subcategories and we will not consider them further. Second, three degrees of uncertainty are distinguished: *technical uncertainties* (data quality, appropriateness of data), *methodological uncertainties* (choices between conflicting data, defining causal relationships between data), and *epistemological uncertainties* (model limitations, indeterminacies, they include lack of knowledge, subjective, and informative uncertainty) and we will devote our attention to technical and methodological uncertainties. There are a number of theoretical frameworks for understanding data quality to decrease their imprecision and uncertainty. One framework is based in semiotics to evaluate the quality of the form, meaning and use of the primary data (Price, Shanks, 2004). One highly theoretical approach analyzes the ontological nature of environmental information systems to define data quality rigorously (Wand, Wang, 1996). The quality of information obtained from processing primary environmental data and information is commonly measured by a number of factors including: *relevance* (does the concept measured corresponds to the concept required); *timeliness* (the period between the time of the observations and the time used); *accuracy* (the deviation between the target value determined by a perfect process and the value determined by the imperfect

[1] Ministry for Environment of the Czech Republic, Praha, Czech Republic,
e-mail: Miloslav_Hlavacek@env.cz, Internet: www.env.cz

[2] Faculty of Informatics, Masaryk University Brno, Czech Republic,
e-mail: 3148@mail.muni.cz, Internet: www.fi.muni.cz

[3] Institute of Biostatistics and Analysis, Faculty of Medicine and Faculty of Science, Masaryk University Brno, Czech Republic, e-mail: hrebicek@cba.muni.cz, Internet: www.cba.muni.cz

process); *accessibility* and *clarity of the information*; *comparability of the statistics*; and *coherence* (Depoutot, 1998), (Statistics Canada, 1998).

The important task of eGovernment Services in current environmental management and decision processes of sustainable development of European Union (including Visegrad countries) is to deal with the primary data uncertainty and using knowledge about data quality to assure that environmental policy and strategic decisions are enough correct and provide added value for improvement sustainable development of European Union (EU). This will reduce the risk of wrong decision in environmental management with respect to minimization of negative environmental impacts and continuous improvement of the state of environment in whole Europe.

European "*Leadership Expert Group*" (LEG) on Quality has been developing proposals for future data quality management in the *European Statistical System* (ESS) (Eurostat News 2002). In order to guarantee the quality of European Statistics, the ESS adopted, in 2005, the *European Statistics Code of Practice*[4] with 15 modified principles. Among the formal definitions that currently exist, the Group considered the International Organization for Standardization (ISO – http://www.iso.org) approach to be must appropriate. There was European AMRADS project, which solved the transfer of identified current best practice methods from centres of excellence to meet users' needs (Charlton, Bailey, 2002), also with respect to data quality and its uncertainty and published *Quality Declaration of ESS*[5] (further *Quality Declaration*). In late 2006 Eurostat has started to develop the overall concept of Data Centres within the framework of Go4 (the European Environment Agency - EEA, DG Environment, Eurostat and Joint Research Centre – JRC ISPRA). As a first step Eurostat has launched a tender in October 2006 about *Implementation of Environmental Data Centres*. It is a specific condition in the tender that the work should be done in close cooperation with Eurostat, other services of the European Commission in Brussels, JRC and the EEA.

Since its establishment in 1997, the Topic Centres of EEA have supported DG Environment in collecting and processing data reported by the Member States of EU in pursuance of the *Standardised Reporting Directive* (Directive 91/692/EEC). This work is expected to be done next years by the *Data Centres* and also its part will be devoted to solving primary data uncertainty and using knowledge about data quality to assure the correct environmental decision.

For example, the implementation plan 2007 for the *European Topic Centre on Resource and Waste Management* of EEA has tasks devoted to waste data management, which the Topic Centre will carry out in 2007 (Mortensen, 2006). The task 7.1.2 Data Centres on waste, resources and products followed from the request of Eurostat to fulfil the *Regulation 2150/2002/EC, on waste statistics* (Regulation No 2150/2002) and *Regulation 574/2004/EC amending Annexes I and III to Regulation 2150/2002/EC* (further *Regulations*).

The role of uncertainty analysis in U.S. Environmental Protection Agency (EPA) has been increased (Krupnick at al, 2006) since the end of the last century. Uncertainty is classified as one of four types: *parameter uncertainty, model uncertainty* (that is, does the model accurately represent and simulate conditions that may exist at a waste disposal site?), *decision rule uncertainty* and *variability*. EPA uses its *Quality System* to manage the quality of its environmental data collection, generation, and use (http://www.epa.gov/quality/). The primary goal of the *Quality System* is to ensure that our environmental data are of sufficient quantity and quality to support the data's intended use. Under the EPA Quality System, EPA organizations develop and implement supporting quality systems.

Further primary environmental data are discussed with respect their quality.

[4] http://epp.eurostat.ec.europa.eu/pls/portal/docs/PAGE/PGP_DS_QUALITY/TAB47141301/CODE_OF_PRACTICE-EN.PDF

[5] http://amrads.jrc.it/WPs%20pages/Quality/quality.htm

2. Primary Environmental Data

Collected primary environmental data is the treasure which helps decision makers, experts and even the public to be aware of environmental situation of the Earth. But this treasure is often in bad condition and quantity itself is not the right key to wealth.

Primary environmental data are monitored and collected in various ways and their quality is also very variable (depending on many factors). These primary data are evaluated and they form required environmental information. If we want to determine the value of such information, it is necessary to measure the quality of the primary data (e.g. by means of *EPA Quality System, Regulations or Quality Declaration*) or complementary by some new methods, which use models of primary data behaviour. In some cases it is necessary to make changes in the primary data (*confirm, add, change* or *delete* their values). When decision makers are making evaluations, conclusions and the decisions from collected/monitored data and processed information from them, they have to assess data and information quality not to make a fatal mistake in the qualified decision.

Decision about the quality of the primary data can be made in various manners. Very often it is not made or used at all and this is the case, which is suitable for some alternate approach. The data quality is (or it can be) judged by less or more experienced authority sometimes, but this process is time and money or other resources consuming and thus not suitable for mass data analysis – this is the reason, why it's not often accomplished. Quality Declaration, Regulations and EPA Quality Systems approaches are not useful at this case as there is no good knowledge about the system of data collection, which is crucial with the two approaches. Alternate approach is welcome also in the case when human factor often comes into play. It is necessary to use some techniques, which will be suitable for automatic processing (processing without the intervention of some person). To compute the quality or even to replace missing and incorrect primary data we can use different models. This is the right direction for automatic processing.

And even worse – there is also no well known and defined standard of the model of primary environmental data behaviour. The present work will introduce one such prototype of model – easy to implement, covering all basic needs in environmental informatics and promising easy model knowledge sharing – it will be suitable to make a basis for appropriate model library.

Paper also discusses the use of the model and its role in the processes of measurement of data quality (Quality Declaration, Regulations, EPA Quality System). Short conclusion suggests future exploitation of possible new ways of dealing with the primary data uncertainty and marks few threats, which can delay the process.

3. Proposed Model of Data Quality Assessment for EGovernment Services

The primary environmental data management is often simulated by the model, but there is no well known and accepted standard of the model. Most of model's experts use their own implementations and their own concept of the model, which have some common features.

The main strength of the new proposed model prototype lies in using ICT for representation of the data and easy implementation, and sharing of the model' results. The purpose of the model is to catch the data uncertainty arising mainly from human factor and also to catch the data uncertainty when any other ways of uncertainty measurement are not present.

First, we have to define all the terms and assumptions which will be used in the proposed model concept implementation. These definitions and assumptions form the model concept.

The primary data sets (databases) are the real input of the model. These data can be characterized by the data model (e.g. Entity Relation Diagram). This data model contains all the dependencies of the entity attributes and relation tables. One "*element*" of the data set (it will be called an item) is formed by its metadata including attribute name, further its value and its key to the rest of database.

Enhanced primary data set is the data set, which contains tags for every its item. The present model makes use of two tags: *the first tag represents the probability of item value to be correct* and *the second* *tag the probability of the item value to be useful for users*. These are two different tags, because the value can be false, but we know that it is close to the true value and on the opposite hand we can have the true value which is, for some reason, completely useless for us.

The data can be accompanied by more tags, because they can have additional important characteristics (measurement device accuracy, etc.), but proposed model uses only the two earlier mentioned ones (probability and usefulness). This decision is motivated by practice and also by the purpose of the model – to catch mainly those errors produced by human factor (or similar factors, e.g. wrong organisation or complete failure of measuring device, etc.).

It is supposed to react only on some kinds of data uncertainty (as mentioned above). There are many sources of uncertainty, including: *uncertainty in scientific constants, observation error, implementation uncertainty*, etc. but it is supposed they can be solved separately by other models or by *EPA Quality System* means and they can be later incorporated into new model (or vice-versa).

The model is defined as a parametric function of more independent variables. Often this is very complicated function (with a lot of exclusions), but not always – sometimes it can be simple. Function value represents item's new value. Input parameters of the function include various knowledge about the item, the value of the item itself and values of the item tags. Models can be easily combined as functions do the same.

When dealing with the data, we have to fill the values of probability and usefulness for all items and this is the new approach. It can be done very simply (by setting some default values) or by application of some rules (e.g. all the data from some sources are more suspicious of being wrong – that means setting their probability lower than the others). Application of rules may be cumulative and this implies the need for some arithmetic means to compute the final value of probability or usefulness.

The last application of the rules would be the comparison with the model of primary data management. If the value of the item is not far from the value suggested by the model, the probability of the item value to be true is high, similar rules apply for usefulness.

Finally, we will get the data with some new attributes and we can decide what to do subsequently – we can define some rules for this. Either we can replace item values with low probability by the output model values or we can exploit some values with high usefulness.

4. Proposed Approach for Visegrad Countries

In previous parts we analyzed the state-of-art data quality assessment for eGovernment Services and also proposed some new possibilities of data quality assessment. The question now is which approach to choose. The context of the question should be regarded from the point of view of Visegrad countries (V4). These countries (the Czech and Slovak Republic, Hungary and Poland) are Members States of EU and they are still located on the crossroad of data quality assessment and will have to choose the right approach to fulfil all needs of legislation and still with the low financial and human-resources costs.

The null-variant of approach (when the data are used as they are, without any or with uselessly low uncertainty assessment or treatment) is always available, of course. To be honest this approach is used more often in V4 countries than it should be used. The reasons lie mostly in the lack of the research, knowledge, human-resources, finances and equipment.

The approach looks to be the best choice from the financial point of view, but it is not always true. The costs which arise in case of wrong decisions based on the inequality data might easily overgrow the cost economies given by not using of proper uncertainty treatment.

OECD countries like USA and some developed Western European countries can use the full approach, which employs the latest knowledge and research of EPA, EEA, Eurostat or Go4 and V4 national authorities. These countries will be able to finance all necessary parts of the approach.

They do not solve the financial compromise between the effort and the effect, as they want to have the best knowledge about the environment.

The typical situation for the V4 countries trying to use this approach is the failure. The principles working in some developed OECD countries are not often working in V4 (from many more or less objective reasons). The resources needed to customize the solution are not available, primary environmental data are often not complete and the human-resources are not well trained. The promised results are not good enough and the failure ends with the frustration and the suspect of the future use of this advanced methods.

Our presented new approach for V4 enables the use of a financial compromise in eGovernment Services. We know there is a need to invest the resources in eGovernment Services wisely to get the best results from the primary data as possible. The first issue is the need to comply the legislation. Then the approach makes use of only the existing data sources and their evaluation is made in the reasonable extent. Final mapping of all evaluated factors into two main attributes (probability and usefulness) ensures easy usability of the data for their evaluation with respect to uncertainty.

The only drawback of the approach is the need for experienced evaluation expert. This need is partially fulfilled by this paper. Also some case studies have been finished, including (Hřebíček, Hejč, 2006), where is shown that even 30% overall waste production clarification can be done with a very limited resources.

The approaches are enough self-explaining – null, complete or the compromise? It is clear that the null approach shouldn't be used any more – either from the point of legislation or the possible financial problems in case of wrong decision. Complete approach is the desired one, but not always available in V4 countries. New presented modular approach in Annex A gives the V4 countries the chance of complying the legislation and the practice with the payable costs.

5. Example of Comparison of Approaches

The differences in approaches are best shown by some specific example. We choose the example of the household waste production at municipalities. In V4, the waste producing by citizens at home are collected and separated into the waste containers depending on the collection system at given municipality. The amount of this waste production and disposal in the given municipality is announced / reported in compliance with the national legislation of the Czech Republic to Ministry of Environment through local state administration bodies and Centre of Waste Management (http://ceho.vuv.cz/). All available annual reports are evaluated and overall production of municipal household waste is aggregated into the final information, which is presented in the environmental reports of the Czech Republic.

This is the common part for all mentioned approaches – the primary data about the municipal household waste production are collected and evaluated. But there are differences in the types of data collection and their processing. When the null-variant comes into play, annual reports of municipalities are just collected and the plain summary is processed and evaluated. Sometimes some most flashy cases of errors are filtered (by means of interval arithmetic).

But we know more about the nature of these data. The municipal household waste production strongly depends on the number of inhabitants and the standard of living. Then there are some other dependencies on size of the community, type of housing, unemployment rate, etc. All these dependencies can be incorporated into the model of behavior of these primary data. Such model forms the knowledge and can be used for verification of the data or to replace the gaps of the data.

Full variant would employ the standardized measuring procedures in the process of primary data collection. Their results would be carried into pre-prepared models of waste production by standardized ways, based on verified knowledge. This would case the validation of the data, which as standard-shaped results would be passed to evaluation process. The results of such (full) approach are much more valuable then the results of previous approach. This approach is used by EPA Quality System, for example.

The compromise approach is often used by statistic offices. But this type of compromise used by statistic offices is very weak. Statistic offices sacrifice the quantity of primary data for their quality in this case. That means only the reports from few towns and communities are collected and verified. The rest of the data (or the huge gaps of the data in other words) is presumed by the model, based on those few collected data. But this approach is "weak", because it doesn't make a use of all available data for the model. For example the Czech Statistics Office doesn't make use of the housing type information in its model, the size of the community-dependency is not continuous but in steps, etc.

The use of the compromise should be as "strong" as possible (as our new approach suggest). It should use all the knowledge which is already available. That means all available reports (no matter if verified or not), all available demographic and geographic data (economic activity, type of housing, age, etc.) and also all the knowledge from the past. Then the resulting verification is more complex, but this is no loss, because more complex automated process is still automated for its user and better results are getting with the same human workload. This compromise approach is then even "stronger" then the full approach, because it has to offer some knowledge to overcome the missing procedures in the data collection process, elimination of human factor influence being the most significant of them.

There is formally described simple model of waste production function

$$P = F(\, \#inh,\ spec,\ std,\ sz,\ unemp,\ hsg,\ heat\,),$$

where are defined:

P	Waste production per year;
#inh	Number of inhabitants;
spec	specific waste production coefficient (reference values of other coefficients) [kg];
std	standard of living coefficient;
sz	size of the community coefficient;
unemp	unemployment rate coefficient;
hsg	type of housing (recreation, blocks of flats, empty houses…) coefficient;
heat	type of heating coefficient.

For example, thus simple function can be defined as (Figure 1):

$$P = \#inh * spec * std * sz * unemp * hsg * heat \,/\, 1000\ [t].$$

Further we can consider coefficients*:* $x = (act \,/\, ref)$, where

ref	reference value;
act	actual value;
x	one of above mentioned coefficients;
cx	compensator of the respective coefficient x (given by optimization process).

Model of standard of living value as one example of numerous sub-models is used to compute actual and reference value of the respective coefficient: $stdV = Rinc*Rsz$, where:

stdV	standard of living value;
Rinc	average income in the given region;
Rsz	size of the community in region coefficient.

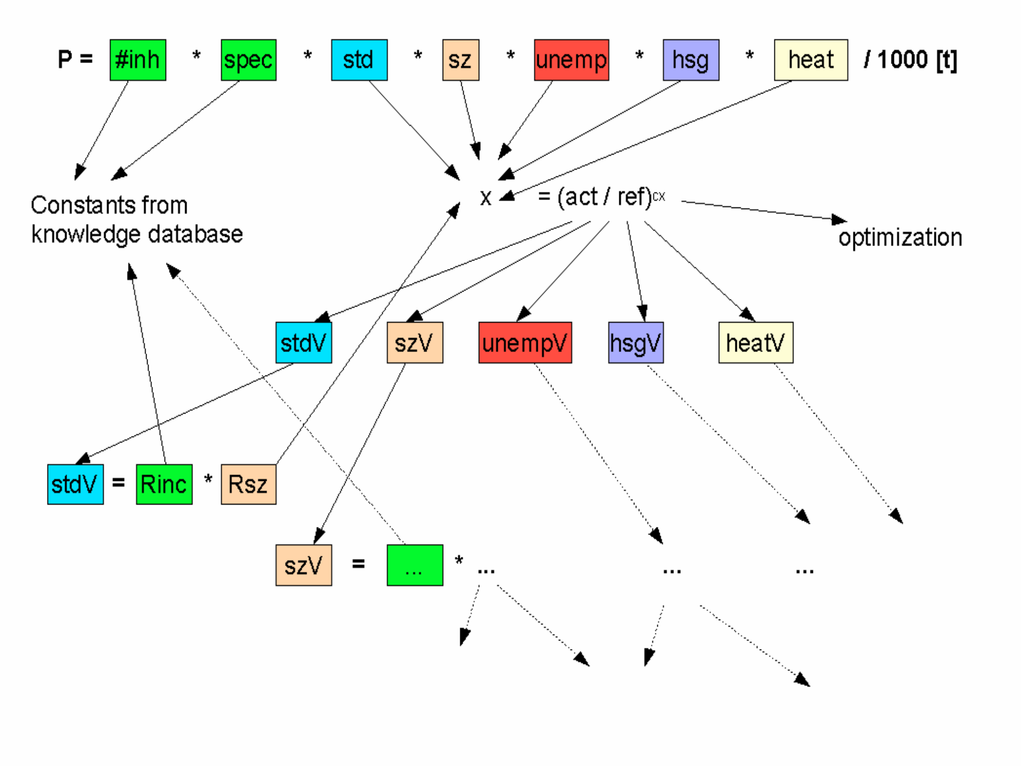

Figure 1: Model of waste production in municipalities

5. Conclusions

Standard approach of dealing with the data uncertainties in Data Centres of the EEA is burdened by inherent limitations – difference in quality of various data sources results in inconsistency of uncertainty treatment and thus in the low quality of information and environmental indicators generated by the data. It is not possible to easily overcome the problem without application of some new approaches.

The new developed approach for eGoernment Services is currently explored and used in parts. Expected benefits comply with the Go4 expectation and they offer significant improvement of the environmental information quality, carried by the environmental data. This means finance savings and less environmental risk in our case. As the process of data quality assessment in eGoernment Services becomes more automated, the new approach can generate even less work load or the same work load with better results to make available quality environmental information and disseminate it to the public. But the exploration is far to be complete at this moment and there is still much to do – especially in the field of particular waste modules. But it could be used in Data Centres of the EEA.

We also show the difference in approaches from the point of view of V4 countries and compares them. “Null” approach was also discussed as the specific problem of V4.

Acknowledgment

The paper is supported by the project Communication Framework No. SM 99/10/05 of Ministry for Environment of the Czech Republic and the DEMO-net project of the European Commission.

Bibliography

Eurostat News (2002): Quality in the European statistical system – the way forward. Luxembourg: Office for Official Publications of the. European Communities.

Charlton, J. and Bailey, S. (2002): Sharing Best Methods and Know-How for Improving Data Quality, http://amrads.jrc.it/WPs%20pages/Quality/quality.htm

Depoutot, R., (1998): Quality of International Statistics: Comparability and Coherence Proc. Conference on Methodological Issues in Official Statistics, Stockholm.

Hřebíček, J., Hejč, M. (2006): Annual assessment of Waste Management Indicators in South Moravia Region (in Czech), Report of South Moravia Government Brno.

Krupnick, A., Morgenstern, R., Batz, Nelson, P., Burtraw, D., Shih, J. and McWilliams, M., (2006): Not a Sure Thing: Making Regulatory Choices under Uncertainty.

Mortensen, L., (2006): ETC/RWM Implementation Plan 2007. European Environment Agency, 15.

Price, R. and Shanks, G., (2004): A Semiotic Information Quality Framework, Proc. IFIP International Conference on Decision Support Systems (DSS2004): Decision Support in an Uncertain and Complex World, Prato.

Regulation (EC) No 2150/2002 of the European Parliament and of the Council on waste statistics, http://europa.eu.int/eur-lex/pri/en/oj/dat/2002/l_332/l_33220021209en00010036.pdf

Statistics Canada, (1998): Quality Guidelines, Third edition, Ottawa.

Van Asselt, M., (1999): Uncertainty in Decision Support: From Problem to Challenge. Maastricht, Netherlands: University of Maastricht, International Centre for Integrative Studies.

Wand, Y. and Wang, R., (1996): Anchoring Data Quality Dimensions in Ontological Foundations, Communications of the ACM, November 1996, 86-95.