GEODESY AND CARTOGRAPHY
Vol. 65, No 1, 2016, pp. 3-12

DOI: 10.1515/geocart-2016-0006

Accuracy assessment of planimetric large-scale map data for decision-making

Adam Doskocz

University of Warmia and Mazury in Olsztyn
The Faculty of Geodesy, Geospatial and Civil Engineering
Department of Land Surveying and Geomatics
12 Heweliusza St. Olsztyn, Poland
e-mail: adam.doskocz@uwm.edu.pl

Received: 14 October 2015 / Accepted: 20 January 2016

Abstract: This paper presents decision-making risk estimation based on planimetric large-scale map data, which are data sets or databases which are useful for creating planimetric maps on scales of 1:5,000 or larger. The studies were conducted on four data sets of large-scale map data. Errors of map data were used for a risk assessment of decision-making about the localization of objects, e.g. for land-use planning in realization of investments. An analysis was performed for a large statistical sample set of shift vectors of control points, which were identified with the position errors of these points (errors of map data).

In this paper, empirical cumulative distribution function models for decision-making risk assessment were established. The established models of the empirical cumulative distribution functions of shift vectors of control points involve polynomial equations. An evaluation of the compatibility degree of the polynomial with empirical data was stated by the convergence coefficient and by the indicator of the mean relative compatibility of model. The application of an empirical cumulative distribution function allows an estimation of the probability of the occurrence of position errors of points in a database. The estimated decision-making risk assessment is represented by the probability of the errors of points stored in the database.

Keywords: models of empirical distribution of errors; accuracy of map data; risk of decision-making

1. Introduction

Digital map data are important for many government departments, business research and the general public. The most accurate are large-scale map data which are produced by geodetic and cartographic procedures. These include, among others: new surveys performed by electronic tacheometers or RTN/RTK GNSS equipment, re-calculation of previous direct measurements and graphical-and-digital processing of analogue maps. Although these procedures are complex, they also provide continuous updating for planimetric data collection. The conducted studies show the position errors of points stored in analyzed databases, which are with mean errors from (0.04 m, 0.14 m) for new surveys and to (0.14 m, 0.46 m) for graphical-and-digital processing of analogue maps (Doskocz, 2013).

However, there are also outliers and gross errors which distort the error distributions. This is in accordance with the law of the propagation of distributions (ISO, 2004), which is a generalization of the law of the propagation of uncertainties (variances).

Therefore, since the application of classical statistical analyses did not produce the expected results (Doskocz, 2005), the author proposed a robust assessment of the accuracy of large-scale digital maps (Doskocz, 2014b) and the application of empirical cumulative distribution functions for a risk assessment of decision-making (presented in this paper).

The paper is not related to natural hazards, for example, of landslides (e.g. Pradhan et al., 2011) or floods (e.g. Hejmanowska, 2006) but refers to an accuracy assessment of map data for decision-making, for example, in detailed localization of objects for land-use planning.

This paper proposes a geomatic analysis for estimation of risk decision-making by data read from the large-scale maps or stored in their databases (using so-called "large-scale map data"). This is an important problem because stored large-scale map data often have incomplete information about data lineage and accuracy as well as erroneous information about position of points and other objects (Doskocz, 2013).

In such situations, a database may mix good data with bad (Siegrist, 2011). In conventional statistical and other type of analyses, bad (erroneous) data are ignored (outliers and gross errors are rejected). Sometimes all data are analyzed by robust methods and by statistical data predictions (e.g. Pita et al., 2011). In the literature was proposed the creation of thematic risk maps using geoinformation systems (Shokin et al., 2011). Guryev et al. (2014) also proposed the construction of dynamic (digital) risk maps whose main goal is to serve as an early diagnostics tool for decision-makers. Potential territorial risk is an assessment of arbitrary points (x, y) located on a metropolitan area map.

In the presented study an analysis was performed for a large statistical sample set of shift vectors of control points, where the shift vector of a control point was identified with the position error of this point. The established models of the empirical cumulative distribution functions of shift vectors of control points involve polynomial equations. The established models of the empirical cumulative distribution functions of shift vectors of control points involve polynomial equations.

Risk analysis is an accurate technique if it fulfills the appropriate criteria. Guidance and a review of the literature are available in Pradhan et al. (2011). An evaluation of the compatibility degree (R^2) of the polynomial with empirical data is presented and an assessment of the accuracy of models by the convergence coefficient:

$$\Phi^2 = \sum_{t=1}^{n} E_t^2 \Big/ \sum_{t=1}^{n} \left(y_t - \bar{y}\right)^2 \tag{1}$$

computed for several-element populations of remainders (there were n remainders of non-linear model E_t determined with a step of 0.05 m within the limits of the scattering of elements of a given set) is also given. The compatibility characteristics of non-linear models with empirical data was represented by

$$quasiR^2 = 1 - \Phi^2 \tag{2}$$

and the indicator of the mean relative compatibility of model

$$\Psi = \frac{1}{n} \sum_{t=1}^{n} \left|E_t\right| \big/ \left|y_t\right| . \tag{3}$$

In the formulas y_t is the size of a variable predicted by the model and $\bar{y}$ is the mean value of the real size of a variable (Gładysz and Mercik, 2007).

Research using the classical statistical method of large-scale map data accuracy assessment including by an analysis of errors distribution did not produce good results.

This paper presents a new approach for accuracy assessment of planimetric large-scale map data by using empirical cumulative distribution functions of position errors of control points.

2. The study data sets

The studied data sets include large-scale map data which may be characterized by their main production method (Dąbrowski and Doskocz, 2008). An analysis was performed for sets of control points of the 1st accuracy group (the so-called "well-defined" points, which comprise three types of point objects: apex points of building contours, boundary points of parcels and points corresponding to above-ground technical utilities).

It should not be forgotten, however, that the earlier-created databases are updated primarily by new survey methods. The diversity of the situational data acquisition methods for large-scale maps is given in Figure 1.

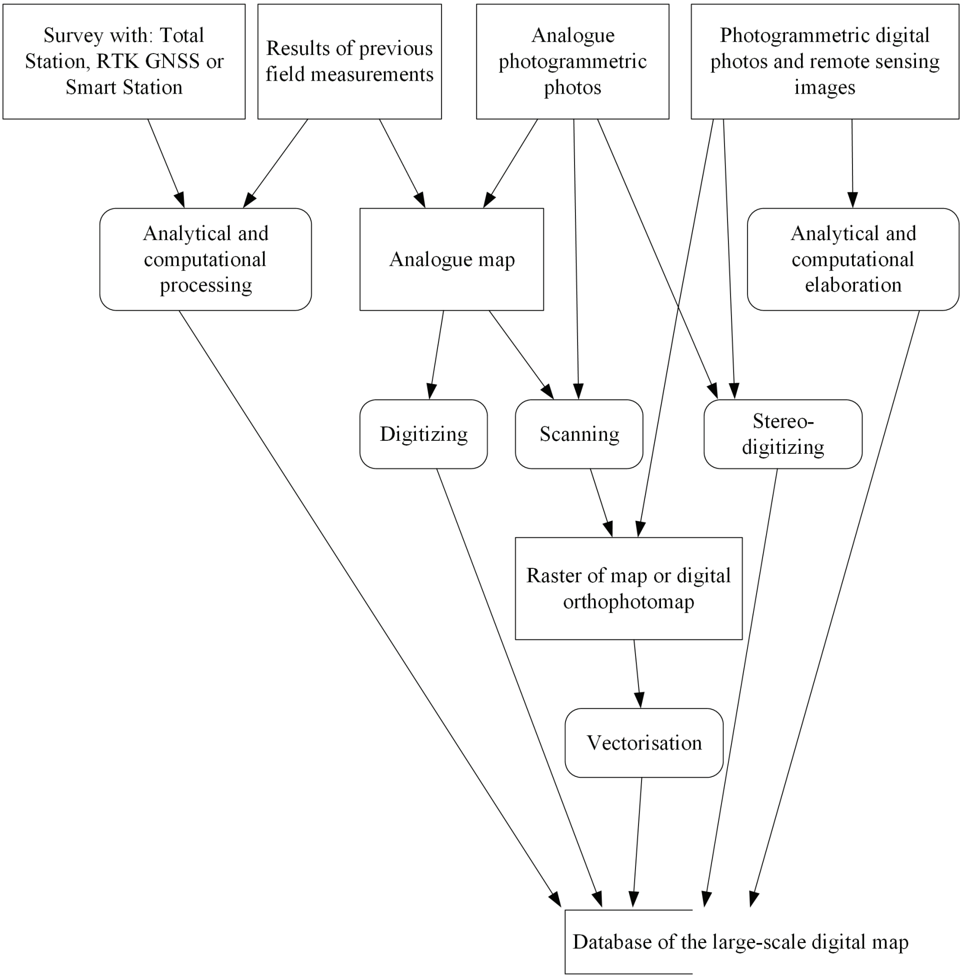
Survey with: Total Station, RTK GNSS or Smart Station
Results of previous field measurements
Analogue photogrammetric photos
Photogrammetric digital photos and remote sensing images
Analytical and computational processing
Analogue map
Analytical and computational elaboration
Digitizing
Scanning
Stereo-digitizing
Raster of map or digital orthophotomap
Vectorisation
Database of the large-scale digital map

Fig. 1. Methods of the situational data acquisition for large-scale maps (Doskocz, 2013)

The studies were conducted on four data sets. Large-scale map data set A was produced based on a direct survey performed with use of an electronic tacheometer and 484 control points were used to assess risk decision-making. Large-scale map data set B was produced based on surveys, performed from 1974 to 1999 using the orthogonal measurement method and, in recent years, by the polar method using an electronic tacheometer. 1,636 control points were used for the risk decision-making assessment. Large-scale map data set C was acquired by manual vectorization of the raster image of an orthophotomap. 773 control points were used to assess risk decision-making. Large-scale map data set D was produced using a graphical-and-digital processing method (by vectorization) of the analogue base maps on a scale of 1:500 with the layers of utilities, on scales of 1:500 and 1:1000. 2,287 control points were used to assess risk decision-making.

The analysis was performed for large statistical samples sets of shift vectors (*V*) of control points.

$$V = \sqrt{dX^2 + dY^2} \tag{4}$$

where *dX* and *dY* are their components, i.e. differences of coordinates.

In the case of map data A, produced using a direct survey with an electronic tacheometer, the coordinate differences were represented by differences between positions of the same control point obtained from two separate tacheometer measurement stations. In the case of other methods of large-scale map data collection, the coordinate differences were represented by differences of coordinates acquired from the investigated data sets and coordinates calculated from control direct surveys. The control surveying was realized (Total Station) by double measurements of a single-signaled point of the 1st accuracy group, which allowed an evaluation of their accuracy (0.01-0.03 m) ± 0.01 m. Thus, the shift vector of the control point may be identified with the position error of this control point.

3. Results and discussion

3.1. Empirical cumulative distribution functions

The empirical cumulative distribution function for decision-making risk assessment presented in this paper was determined using an evaluation of the compatibility of the model with empirical data. In Figure 2 was presented on vertical axes probability that error not exceed the size, which was presented on horizontal axes (in meters).

The main evaluator is the coefficient of determination (R^2), containing information on the percentage variation of the empirical set elements which are explained using the formulated analytical model. This coefficient takes values from the interval (0, 1). A higher R^2 indicates a better fit of the model to the data and, in practice, should aim for the highest determination coefficient.

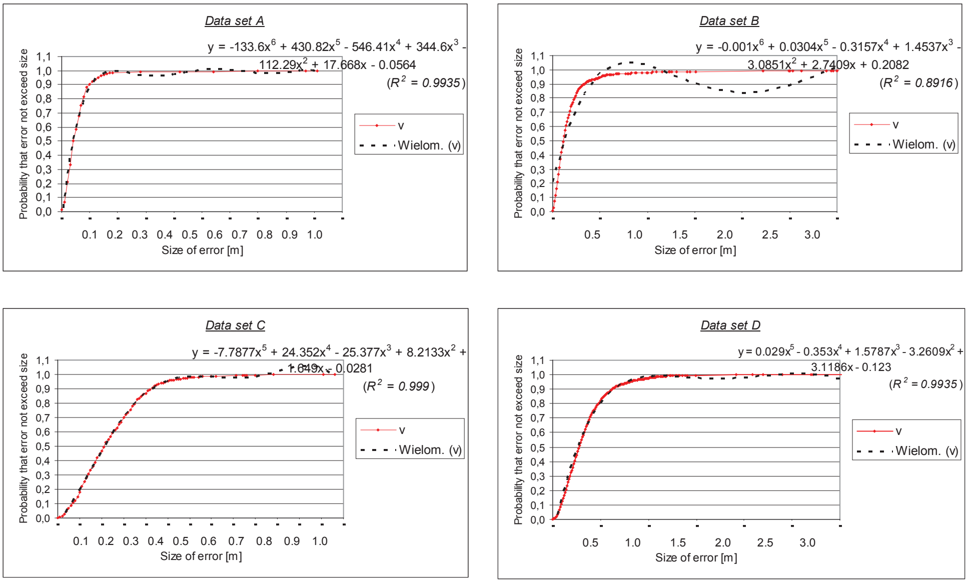

Fig. 2. The empirical cumulative distribution functions characteristic for data sets and matched to them analytical models (polynomials)

The polynomial equations models of the empirical cumulative distribution functions of shift vectors of control points in studied data sets, with a maximum vector value (v_{max}) and number of set elements (N) were stated, and are given in Table 1.

Table 1. Polynomial equations which are models of empirical cumulative distribution functions of position errors of points and their assessments

Data set A $y = -133.6x^6 + 430.82x^5 – 546.41x^4 + 344.6x^3 – 112.29x^2 + 17.668x – 0.0564$ (v_{max} = 0.91 m, N = 484, R^2 = *0.9935*) n = 18, Φ^2 = 0.065, *quasi*R^2 = 0.935, Ψ = 0.019	*Data set C* $y = -7.7877x^5 + 24.352x^4 – 25.377x^3 + 8.2133x^2 + 1.649x – 0.0281$ (v_{max} = 1.26 m, N = 773, R^2 = *0.999*) n = 22, Φ^2 = 0.009, *quasi*R^2 = 0.991, Ψ = 0.050
Data set B $y = -0.001x^6 + 0.0304x^5 – 0.3157x^4 + 1.4537x^3 – 3.0851x^2 + 2.7409x + 0.2082$ (v_{max} = 15.87 m, N = 1636, R^2 = *0.8916*) n = 25, Φ^2 = 0.130, *quasi*R^2 = 0.870, Ψ = 0.076	*Data set D* $y = 0.029x^5 – 0.353x^4 + 1.5787x^3 – 3.2609x^2 + 3.1186x – 0.123$ (v_{max} = 4.85 m, N = 2287, R^2 = *0.9935*) n = 25, Φ^2 = 0.009, *quasi*R^2 = 0.991, Ψ = 0.060

The established models of empirical cumulative distribution functions satisfy the accepted correctness criteria (Gładysz and Mercik, 2007): the coefficient of determination (R^2) on a level of not less than 0.6 and an indicator of the mean relative compatibility of the model (Ψ) of not more than 0.1. Accordingly, the models may be used to estimate the probability occurrence of position error of a control point in digital map data sets.

The high probability of the occurrence of large errors in the data set suggests a large risk in decision-making about the localization of objects, e.g. for land-use planning and for other purposes (e.g. Andrew et al., 2015).

3.2. Risk assessment of decision-making based on values of the position errors of points

In the presented results, the exact size of position errors of points provides a risk assessment of decision-making based on stored data sets.

The application of an empirical cumulative distribution function enables the calculation of the probability of the position errors of well-defined points (with sizes depending on the object of interest) in databases. This could be used, for example, to meet the accuracy standards required to determine an object's position (with a position error of control point not greater than 0.10 m) or, for the old boundary points of parcels, the probability that a position error of point does not exceed 0.60 m.

The decision-making risk assessment based on the occurrence of the position errors of points in sets of large-scale map data estimated from empirical cumulative distribution functions is given in Table 2.

Risk assessment of decision-making is possible by the assessment of accuracy of well-defined points, for example, for determining the position of boundary points of the parcels in the use and modernization of land records.

Table 2. Risk assessment of decision-making estimated by the probability of the position errors of points stated from empirical cumulative distribution functions

Data set	The probability of position errors of points not exceeding this size in [m]								
	0.05	0.10	0.15	0.20	0.30	0.40	0.50	0.60	0.70
A	0.67	0.91	0.99	0.99	*0.96*	*0.98*	1.00	1.00	*0.98*
B	0.36	0.47	0.57	0.66	0.80	0.90	0.98	1.00	1.00
C	0.10	0.22	0.36	0.49	0.72	0.88	0.96	0.98	0.98
D	0.05	0.18	0.30	0.40	0.57	0.71	0.81	0.88	0.93

In the Table 2 was presented information (metadata) on the probability of meeting the planimetric accuracy standard for a large-scale map (for maps on the scale 1:500, the position error of well-defined points should not be greater than 0.15 m, 0.0003 m × 500 = 0.15 m) and a robust evaluated database with an occurrence of excessive position error of point (the probability of no occurrence of errors with outlier sizes and gross errors).

For the assessed databases, the reliability of digital map data for the level of the planimetric accuracy standard is satisfied by the data obtained by new surveys performed using an electronic tacheometer (data set A).

The accuracy of large-scale map data produced based on surveys (data set B), performed over 30 years using the orthogonal method of measurements and updated, in recent years, by the polar method using an electronic tacheometer, are also quite satisfactory. More than half of the details of the 1st accuracy group meet the planimetric map accuracy standard.

For the data acquired by manual vectorization of the orthophotomap raster image (data set C), more than one-third of the objects meet the planimetric database accuracy standard on a scale of 1:500.

In the case of data produced using the graphical-and-digital processing method for analogue base maps on a scale of 1:500 (data set D, vectorization preceded by scanning maps), less than one-third of the objects are compatible with the planimetric database accuracy standard.

A credibility evaluation of the analyzed large-scale map data by estimating the probability of no errors with outlier sizes is also interesting (which may be a “confidence index” for database users). The level of errors classified as outliers was established for values greater than 0.68 m (Doskocz, 2013).

Therefore, the position errors of control points at the 0.70 m level reduces the confidence index for the database (Table 2). For data sets A, B or C, verification at the position error of point at the 0.60 m level provides a 100% level for the “confidence index” for these databases. In contrast, the use of data set D lowers the data credibility to the level of about 90% and increases the decision-making risk based on this data set.

For example, the American standard for spatial data accuracy contains the risk of the unknown accuracy of used data sets (NSSDA, 1999).

4. Conclusion

Large-scale digital maps are an important part of the spatial data infrastructure (SDI), and the registers of land and buildings are especially relevant for the European Community (Directive, 2007) and provide reference data to other SDI resources. In addition, cadastral databases play a key role in the structure of government information (Lewandowicz, 2002; Act, 2010). The accuracy of data stored in cadastral and other government registers determines the quality and credibility of decisions based on geomatics – defined as “computer-assisted decision-making tools” (e.g. LGA, 2014).

In addition, geomatics can be viewed as bridging the gap between the producers of digital map data and the data users (Burkholder, 2005). This topic should be considered for aspects of typical tasks of a large-scale map (Doskocz, 2014a) or land information system (Bielecka and Całka, 2012).

In this paper, it was demonstrated that the established analytical models allow a risk assessment of decision-making based on planimetric large-scale map data. Empirical cumulative distribution functions were used to estimate the probability of the occurrence of position errors of points in data sets and to indicate the risk of decision-making about the localization of objects, e.g. for land-use planning.

The estimated empirical cumulative distribution functions have high correctness (Table 1). However, in the model there may be some deviations from the cumulative distribution function (shown in Figure 2). In the studied error sizes (up to about 4.5σ), this is particularly evident in the case of databases with a high level of planimetric accuracy, e.g. data acquired from field measurements.

Confirmation was obtained for data set A, where the complete credibility of the data was found at a position error of points of approx. 0.5 m (Table 2) and the confidence index of the database decreased slightly to approx. 98%. This situation is acceptable because the indicator of the mean relative compatibility of the model was $\Psi = 0.019$ – according to the assessment of the model compatibility with empirical data (Table 1).

The use of databases to support decision-making is associated with problems because decisions are based on data combined from different sources (Harding, 2013). The current state of the creation and modernization of national geodetic and cartographic resources should be revised (Doskocz, 2015) because the metadata of these databases are important for the creation of the European and Global Spatial Data Infrastructure (Bank, 2004).

Acknowledgment

This work was supported by the University of Warmia and Mazury in Olsztyn under Grant number 28.610.017-300.

References

Act, (2010). Polish Parliament, the Act of March 4, 2010: on Spatial Information Infrastructure. Warsaw, Journal of Laws on 2010 no. 76, item 489.

Andrew, M. E., Wulder M. A., Nelson, T. A. and Coops, N. C. (2015). Spatial data, analysis approaches, and information needs for spatial ecosystem service assessments: a review. *GIScience and Remote Sensing*, 52, 344–373. DOI: http://dx.doi.org/10.1080/15481603.2015.1033809

Bank, E. (2004). Importance of open spatial data infrastructure for data sharing. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, Vol. XXXV-B4, 271–276.

Bielecka, E. and Całka, B. (2012). The analysis of the land exclusions from agricultural and forest production in the rural areas. Proceedings of the Polish Academy of Science *Infrastructure and Ecology of Rural Areas*, No. 2/III/2012, 163–173.

Burkholder, E. F. (2005). Geomatics curriculum design issues. *Surveying and Land Information Science*, Vol. 65, No. 3, 151–157.

Dąbrowski, W. and Doskocz, A. (2008). Estimation of accuracy of the large-scale digital topographic map data. *Proceedings Paper of 7th International Conference on Environmental Engineering*, Vol. 1-3, 1293–1299.

Directive. (2007). European Parliament and of the Council, Directive 2007/2/EC of 14 March 2007 establishing an Infrastructure for Spatial Information in the European Community (INSPIRE). Official Journal of the European Union L108 (50) from 25 April 2007.

Doskocz, A. (2005). The use of statistical analysis for estimation of positional accuracy of large-scale digital maps. *Geodesy and Cartography*, Vol. 54, No 3, 131–150.

Doskocz, A. (2013). Methodology for assessing the accuracy of digital large-scale maps. *Dissertations and Monographs* 193, University of Warmia and Mazury in Olsztyn, Poland.

Doskocz, A. (2014a). About accuracy of analytical determination of areas for cadastre and other purposes. *Proceedings Paper of 9th International Conference on Environmental Engineering*, Vol. II, 673–680. DOI: http://dx.doi.org/10.3846/enviro.2014.203

Doskocz, A. (2014b). Robust assessment of planimetric accuracy of large-scale map data. Unpublished manuscript.

Doskocz, A. (2015). The current state of the creation and modernization of national geodetic and cartographic resources in Poland. *Unpublished manuscript*.

Gładysz, B. and Mercik, J. (2007). *Econometric modeling. Case study*. Publishing House of Wrocław University of Technology.

Guryev, E. S., Poluyan, L. V. and Timashev, S. A. (2014). Construction of dynamic risk maps for large metropolitan areas. *Journal of Risk Analysis and Crisis Response*, Vol. 4, No. 2, 72–76.

Harding, J.L. (2013). Data quality in the integration and analysis of data from multiple sources: some research challenges. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, Vol. XL-2/W1, 59–63.

Hejmanowska, B. (2006). Influence of data quality on modeling of flood zones. *Annals of Geomatics*, Vol. 4, No 1, 145–150.

ISO. (2004). Guide to the expression of uncertainty in measurement – Supplement 1: Numerical methods for the propagation of distributions. *International Organization for Standardization*.

Lewandowicz, E. (2002). The role of cadastre in government information structure. *Proceedings of the Wrocław University of Environmental and Life Sciences*, No. 452, 223–228.

LGA. (2014). Available online at: *http://www.ensg.eu/Laboratory-for-Research-in-Applied-Geomatics*

NSSDA. (1999). Positional accuracy handbook. Using the National Standard for Spatial Data Accuracy to measure and report geographic data quality. Available online at: *http://www.mnplan.state.mn.us/pdf/1999/lmic/nssda_o.pdf*

Pita, G.L., Francis, R., Liu, Y., Mitrani-Reiser, J., Guikema S. and Pinelli, J. P. (2011). Statistical tools for populating/predicting input data of risk analysis models. In: B. Ayyub (ed), *Vulnerability, Uncertainty, and Risk: Analysis, Modeling, and Management*, ASCE, 468–476.

Pradhan, B., Mansor S., Pirasteh S. & Buchroithner M. F. (2011). Landslide hazard and risk analyses at a landslide prone catchment area using statistical based geospatial model. *International Journal of Remote Sensing*, Vol. 32, No. 14, 4075–4087.

Shokin, Y. I., Moskvichev, V. V. and Nicheporchuk, V. V. (2011). Method of assessment of human-induced area risk and creation of risk map using geoinformation systems. In: B. Ayyub (ed), *Vulnerability, Uncertainty, and Risk: Analysis, Modeling, and Management*, ASCE, 442–449.

Siegrist, J. (2011). Mixing good data with bad: how to do it and when you should not. In: B. Ayyub (ed), Vulnerability, Uncertainty, and Risk: Analysis, Modeling, and Management, ASCE, 368–373.