Jul 12th, 9:30 AM - 9:50 AM

Design of decision support tools for the quality assessment of illegal dumping notifications based on crowd-sourced data

Jan Brus
Palacký University Olomouc, jan.brus@upol.cz

Jakub Vrkoč
Palacký University Olomouc, vrkock@gmail.com

Miroslav Kubásek
Masaryk University, Institute of Biostatistics and Analyses, kubasek@iba.muni.cz

International Environmental Modelling and Software Society (iEMSs)
8th International Congress on Environmental Modelling and Software
Toulouse, France, Sabine Sauvage, José-Miguel Sánchez-Pérez, Andrea Rizzoli (Eds.)
http://www.iemss.org/society/index.php/iemss-2016-proceedings

Design of decision support tools for the quality assessment of illegal dumping notifications based on crowd-sourced data

Jan Brus[1], Jakub Vrkoč[2], Miroslav Kubásek[3]
Palacký University Olomouc, Department of Geoinformatics, jan.brus@upol.cz[1], vrkock@gmail.com[2]
Masaryk University, Institute of Biostatistics and Analyses, kubasek@iba.muni.cz[3]

Abstract: Illegal dumping has become an increasingly significant environmental problem throughout the Czech Republic and also in Prague. The waste is disposed in areas such as vacant lots, along roadways, alleys, city and county parks, ravines, construction sites. The challenge for cities dealing with illegal dumping is that it is tough to identify when it will happen, where it will occur, and how much will be dumped. The paper is therefore focused on designing a decision support tool, which uses crowdsourcing data to cope with this problem as the management of illegal dumps can be built upon citizen reports via the internet. Many people possess smartphones with data connection to the Internet and are not only limited to Wi-Fi networks. Cellular data transfer is one of the most important prerequisites for the usage of mobile applications which are a huge source of spatial data based on crowdsourcing principles. As the quality of reports in time varies, it is important to focus on the spatial accuracy of reports to increase the efficiency of reporting. For this reason, the spatial quality assessment approach is applied in the application for illegal dumping notifications – "ZmapujTo" for the city of Prague. The model uses the maximum entropy approach combined with several other methods for data validation to automatically access the quality of report based on spatial location.

***Keywords**:* crowdsourcing, Maxent, illegal dump, ZmapujTo

1 INTRODUCTION

Illegal dumping means waste dumping into places, which are not meant for it. These places are mostly readily accessible, and there is a small risk that someone will be seen while dumping illegally. Illegal dumping also covers placing waste in the vicinity of other kinds of dustbins that are not meant for it.

Issues of illegal dumping in Czech waste management have been discussed for many years. The essential reason for the discussion is problematic. In most cases is almost impossible to identify the subject which has established such illegal dump. The subject can be sanctioned for the establishment of such an illegal dump as illegal dumping is an offense. A physical person can be fined from CZK 3 000 up to CZK 50 000. Businesses can be fined up to CZK 10 million. Fines for illegal disposal of waste differs according to the originator.

The current legal situation, Act 185/2001 Coll. on waste and change of other acts, as amended by later regulations, does not fully regulate this problem and regrettably means for a much less effective solution of the removal of illegal dumps than previous legal regulation, Act 125/1997 Coll. on waste. This law, apart from the actually legally binding and effective law, addressed the question of illegal dumps particularly in such a way, that responsibility for an illegally gathered dump on certain land is transferred to the landowner.

Certain hope in a stated field can be seen in the newly prepared act on waste, which returns to the original idea of Act 125/1997 Coll. The law enables municipalities with extended obligations to retain more authority to dispose of illegal dumps and also binds the landowner to take proper care of his/her property. From the point of view of the removal of illegal dumps, one of the possibilities is to use crowdsourcing for the notification of illegal dumps.

1.1 Crowdsourcing

Crowdsourcing is formed from two words: crowd and outsourcing. Crowdsourcing means to hire the crowd to solve a certain problem caused by a crowd, mostly by a bigger group of people, usually over the internet and anonymously (Howe, 2008). The crowdsourcing method utilises efficient communication, interest, and drives the community to create content. The community can compete with professionals mostly through the cost of ownership. The task itself is commissioned to the community, not other organisations or businesses. Advantages of crowdsourcing when collecting spatial data rest in the high volume of newly created spatial data for minimal costs. The reason for this is the cooperation of a higher number of persons than an organisation could afford. One of the main disadvantages of crowdsourcing is the complicated assessment of data quality. Information retrieval is a common part of human society. Mobile devices can be used as a means for fast, cheap, and efficient data retrieval. There is a large number of applications functioning on this principle. The most famous are OpenStreetMap, Waze, SyriaTracker, or Flickr.

1.2 Data quality and crowdsourcing

Data quality is a key subdiscipline of geoinformatics. Data quality is determined not only by basic data, but mostly by the user itself. High-quality data can be marked as "accurate and trustworthy". Nevertheless, this definition and the whole subdiscipline faces an inconsistency of used definitions and inaccurate terminology. Information on the quality of observed data is necessary for their selection and proper use. The aim of data suppliers is primarily their repeated use. For the fulfilment of these conditions, it is necessary to follow users' demands during production or spatial data transformation. The second aspect is identical with documentation supplied to the data users (Brus, 2013). From the point of view of crowdsourcing data quality, there is a very thin line between useful and unusable data. Regular data are standardised as a part of the process. For crowdsourced data, it is very difficult to assess geodata quality without a prior standardisation process. There is a large number of standards for geospatial data. These standards should be implemented within crowdsourcing data creation.

Hans-Jörg Stark from the Centre of Geoinformatics in Salzburg states that the ISO 19100 norm series is suitable for the assessment of the quality of crowdsourcing data (Stark, 2010). These norms were created under ISO/TC 211 and developed standards for the use and search for geographical information. It not only addresses technical and semantical issues, but also metadata which is critical for the assessment of data quality. In some cases, there is no need for high-quality data. Emergency situation mapping can be set as an example (earthquake, floods, fires), where a large number of contributions must be processed at the expense of accuracy. Opposed to companies busying themselves with map making, projects based on crowdsourcing are limited by the cooperation of people within the community (Stark, 2010). Some applications were open for users and were prone to error and vandalism. Certain measures can contribute to higher quality of contributions. A request for user login for editing the main elements can be stated as an example. This "limitation" can lead to an opposite effect; it can lower the number of contributors. Overall, data quality assessment is performed with a lower interest of an expert company.

Jeffrey Heer and Michael Bostock from the Stanford University published a study in 2010 (Heer and Bostock, 2010), where they pointed out that data quality of regular spatial data is well solved. For crowdsourcing data, the situation is different since the data are based on voluntary activity. No one is

responsible for metadata creation. These metadata are usually missing, so a check is necessary. This is an irreplaceable prerequisite for a further analysis and usage of data. One of the possible solutions on how to fill the gap between the usability of data and demanding control of quality is the use of a graphical view of data quality, a so-called visual. The principle is to graphically visualise factors that affect the quality of spatial data.

Matthew T. Rice (Rice et al., 2012) from George Mason University in 2012 published “Crowdsourced Geospatial Data”. Many application users using crowdsourcing are not educated in geographics or geoinformatics. These are denoted as "non-geographers". Non-geographers contribute to the community by mining geospatial data, application development, and other media activities. The usage of geographical techniques and tools for personal and community activities or by a non-expert group of users led to the creation of so-called VGI (Volunteered Geographic Information) as a part of UGC (User Generated Geographic Content). According to the author, the biggest advantage of crowdsourcing is in fast data creation, which he proves on many disasters (earthquakes, floods, fires). Michael F. Goodchild (Sui et al., 2013) addressed the quality of crowdsourced spatial data. He has addressed traditional assessment for several decades. According to him, the most important general tool for assessing spatial data quality are metadata and the maximum utilisation of their potential. According to the author, quality data creation is ensured by the large community of contributors. This is the underlying assumption for the long-term lifetime of projects such as OpenStreetMap. Large companies such as OSM or Wikipedia have developed analytical tools for the detection of malicious contributions. The informed public begins to recognise this type of geodata in a full range. Still, questions are asked about how reliable or trustworthy a given data source is. Jingfeng Xia (Xia, 2012) from the University of Indianapolis describes two basic attitudes in quality assessment. These are quantitative metrics based on objective measurement and qualitative metrics based on subjective measurement. The optimum solution is the connection of both attitudes into a so-called multi-level assessment. M. van Exel (Exel et al., 2010) established research of spatial data quality for the assessment of three core indicators. These are the spatial activity of the community (number of editions in a given place), temporal activity (number of editions in a particular time span), and relative activity of the crowd (number of editions when compared to neighbouring areas). However, for this attitude it is necessary to possess historical information from crowdsourced data..

For working with crowdsourcing data is important to develop a conceptual model for crowdsourcing data quality assessment (Idris et al., 2014). Authors described a method of creating an automated search and comparing processes for their accuracy and subsequent validation. The description of the automated tool is included. This tool uses the conceptual model itself. Authors from from Stanford University have introduced an approach for ensuring quality data using a search, where the margin of error, or inaccuracy during mapping, respectively, is greatly reduced (Sarma et al., 2014). Generalised algorithms for use in some case studies are included. These algorithms are applied to real data. The authors assert that there can be cases of creation and utilisation of crowdsourced data, where maximum spatial accuracy is not required. In this case, the algorithms will be inefficient for use in GIS.

1.3 Application for illegal dumping notifications – ZmapujTo

The first version of ZmapujTo.cz was formed in 2012 as an ecological project for the fight against illegal waste dumping in the Czech Republic and also as a contribution to the solution of the issue of illegal dumping (Kubásek, 2015; Kubásek and Hřebíček, 2013). The project was focused mainly on citizens that do not like the illegal dumping in their towns, municipalities, or the countryside and want to do something about it. Knowing the current situation was the first step. At the time of creation, there was only a database of old ecological burdens, which covered the illegal dumps only marginally. In order to cover the largest possible area and utilise the potential of crowdsourced data, a platform was founded for information gathering from citizens. The modern, efficient, and widely-accepted platform had to be chosen for mapping while using the mobile application and interactive web form for reporting. The first version of the application was very user-friendly and, thanks to the mobile application, any smartphone user could very quickly and easily report illegal dumping. In case the user did not own a smartphone, he/she could report an illegal dump using an interactive web form. More

than 2 500 illegal dumps were reported, and more than 40 municipalities and towns took part during the lifetime of the first version. In March 2014, the second version of ZmapujTo.cz was launched. This version introduced several new features. The goal of the project, to fight illegal waste dumping in the Czech Republic and to make citizens act, remained the same. The most important change was the ability not only to report illegal dumping, but also a variety of other problems that someone can encounter both in town and in the countryside. The entire website was redesigned, including an interactive map for efficient, fast, and intuitive work. Thanks to its responsive design, www.ZmapujTo.cz can be used directly from a tablet or a phone.

1.4 Used technologies

Single regions are sorted in the application according to the ArcČR 500 geodatabase (version 3.1) which is available for free. For reports, Google Maps API is used. Google Street View is also implemented in the application. Therefore, users can scout the given location which makes orientation easy. The Land Register is also available for the detailed preview of the report. The website's responsive web design is based on the Twitter Bootstrap template and is programmed using the AngularJS library as a single-page application with a connection to Google Maps API. The backend server runs on NodeJS (event-managed runtime based on JavaScript) and Express (application framework over NodeJS). For data storage, the document-oriented NoSQL MongoDB database is used. Mobile applications are based on Sencha Touch and Apache Cordova technologies, whose API is built on PHP using Symfony framework, Doctrine ODM. The Mandrill platform is used for sending e-mail notifications.

2 SYSTEM FOR SPATIAL ASSESSMENT OF QUALITY OF INDIVIDUAL REPORTING

The practical part of quality assurance of single reports was created for the capital region of Prague. This area was chosen deliberately as the sequel to the data set supplied in the form of open data by the Planning and Development Institute of Prague. Open data are information and data that can be used without limitations and are available for free. The data are available under CC BY-NC-ND 3.0 CZ (Creative Commons Attribution – Non-Commercial No Derivatives 3.0 Czech Republic license). The license allows copying and further distribution on any media in any format. The condition is keeping the same license. Data published in this way enable further research, analyses, but also commercial use for the creation of internet applications. The Planning and Development Institute of Prague (IPR Praha) is offering geographical data in the form of open data formats and web services. Open data are published in vector formats such as Esri shapefile, geo JSON, GML, DXF, or bitmap TIFF and JPG formats. Apart from bitmap data, open data are published in S-JTSK and WGS-84 coordinate formats. The advantages are similar metadata records. The following data layers were used for the implementation of quality assessment:

- Current land use – WGS 84, 1:5 000,
- Planned land use - (regional plan) – WGS 84, 1:10 000,
- Prague orthophoto – S-JTSK , (pixel 10 cm),
- Bridges 3D – S-JTSK.

Spatial data of the street network of Prague are retrieved from the Register of Territorial Identification, Addresses, and Real Estate (RUIAN). These data are used for optimising the result.

Actual data from the ZmapujTo application are used as a database of illegal dumps. Data are supplied as "Open Data" under license CC BY-NC-SA 4.0 (Creative Commons – Attribution – Non-Commercial-

ShareAlike 4.0 International). The data format is GeoJSON. Data are updated for download with one-day delay from "datacube.io".

2.1 Methods and procedures

In places where illegal dumps were not reported and for the computation of illegal dump probability distribution in places, the maximum entropy - Maxent algorithm for was used. This algorithm was used for modelling the spatial distribution of organisms. The name “Maxent” comes from maximum entropy, i.e. maximum entropy or level of organisation. This model is described by S. Phillips (Merow et al., 2013; Phillips and Dudík, 2008). It is a statistical method, which uses present data. Therefore, all data have to be prepared for use in this modelling software. Data points of illegal dumps in GeoJSON were converted into an Esri shapefile. Input data were selected based on set conditions: a) only illegal dumps, b) middle and large sizes. To make sure that the data were correct and dumps were real dumps and of the correct size, a manual check using photographs was conducted. 686 illegal dumps in the area of Prague were the results of the visual analysis. During another part of the preparation, the SDMtoolbox was used (Brown, 2015). For the removal of data duplicity and influence of neighbouring results, data selection was applied based on spatial autocorrelation with a value of 50 meters. The input layer for analysis finally contained 318 dumps. In the second, data had to be prepared as environmental layers. Spatial data from IPR Praha were processed. Detailed metadata records are available for these data on Prague’s geoportal. For data preparation to be accepted by Maxent, the python based toolbox “PrepareDataforMaxent10_1” toolbox (Dilts, 2015) was used. Attribute records were altered for the subsequent bitmap. Vector data were converted into a bitmap format – the size of one pixel was set to 1x1 meter. Maxent modelling software was used for the single type of occurrence modelling based on environmental variables. A dataset representing places of real occurrence of illegal dumps over 50kg in the area of Prague in the CSV format was used as a first input. ASCII files of the actual situation of land use and functional areas prepared in advance were used as “environmental layers”. The resulting model represents a regular square grid, where every cell (square) represents probability of occurrence of illegal dumps. The resulting bitmap was visually checked using an orthophoto. By manual check, the resulting bitmap was optimised. A layer of current land use was used for optimisation. According to the "KOD" attribute, 66 types of land use with minimum probability of occurrence of illegal dumps were selected by expert estimation (structures, education facilities, etc.). The chosen probability of occurrence of a dump for a given functional area was subsequently integrated into the resulting layer with the goal to lower the probability within this functional area. A solution to possible spatial error when using GPS localisation in the mobile device by setting a possible error to 10m was implemented into the computation at the same time. Infrastructure was optimised after repeated checks of the functional area. This list of infrastructure was gained from RUIAN and converted into the geodatabase. Important structures and streets were exported from the vector layer of the actual situation of land use. Using the “Select by locations” function, infrastructure was used and exported into the new layer. For this infrastructure, a surrounding zone (buffer=3 meters) was computed and the optimisation procedure was repeated in the same way as in the first case. For regular streets, the buffer was set to 2 meters. To ensure the correctness and filtering out the possibilities of "inaccurate" placement of illegal dumps, an optimised bitmap was "smoothed out" using a "Focal statistics" filter (neighbourhood=circle) and radius (cell=10) were set for the computation.

2.2 Results

A raster of probability of illegal dumps in the area of Prague is the result of the whole process of preparation, computation, and optimisation. This result will be used for checking the probability of correctness of citizens' reports since the checks are currently carried out solely manually. For optimisation purposes, the raster was checked from the point of view of prediction accuracy of illegal dumps. In the first place, 52 randomly selected and placed points in the area of Prague were checked. The check was carried out manually over orthophotos by expert estimation. This expert estimation was subsequently observed with values taken over from the optimised raster. The result concluded that 48 points were real and four were misplaced – showing 92% accuracy. In this case, a 13% increase in accuracy in comparison with the non-optimised raster.

In another stage, real data from the ZmapujTo application were used. 314 illegal dumps in Prague were used. These dumps were checked that they really exist or existed in the given places. In this case, there is an expectation that all dumps should be located in places with a higher occurrence probability than the expertly set threshold of 30%. Thirty percent is an expertly set threshold value expressing the probability threshold, where there is a real expectation that an illegal dump really exists. At the same time, it is possible to get the value of the probability of occurrence of an illegal dump at a given place. Direct implementation of the resulting model into the ZmapujTo application is possible in the future. Currently, individual reports are checked manually (Figure 1).

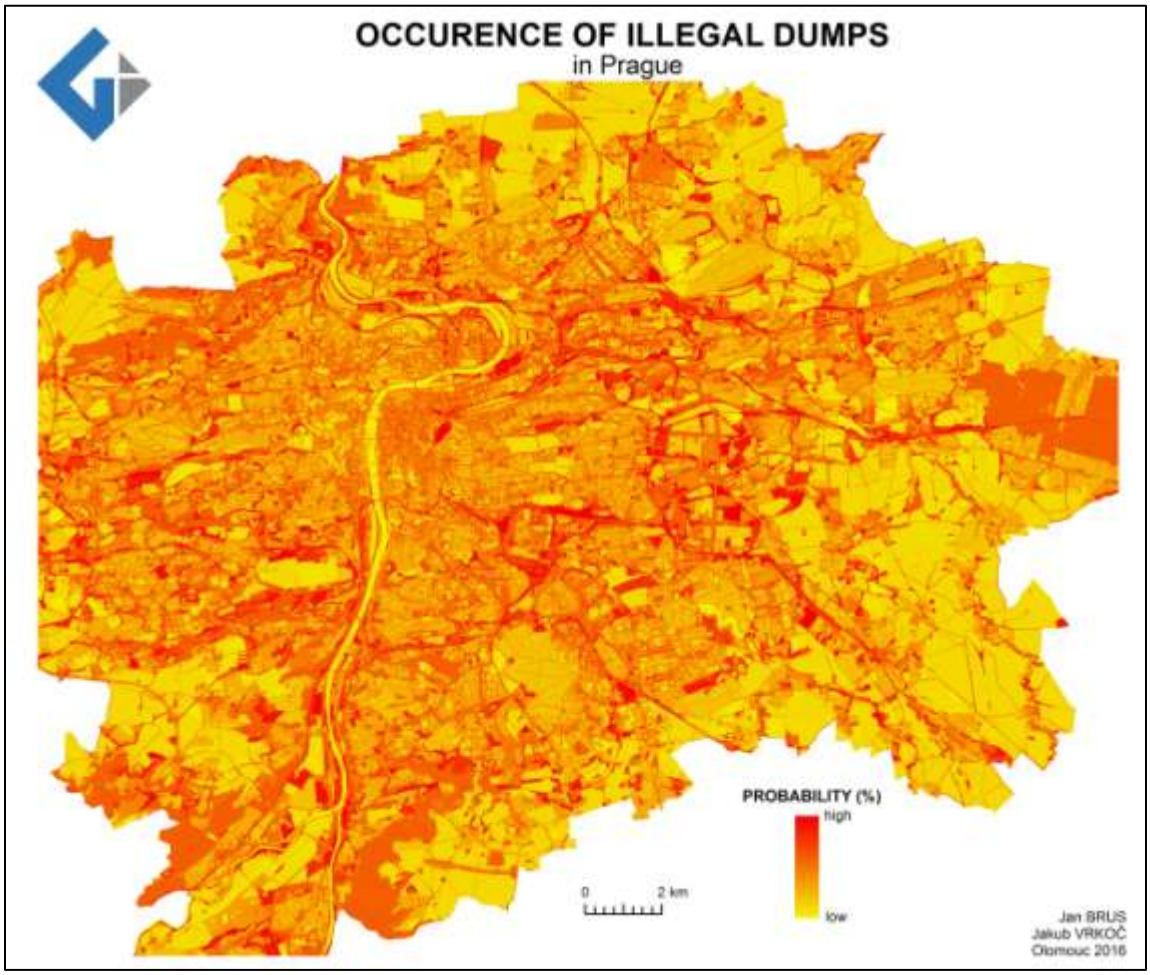

Figure 1: Probability of illegal dumps in Prague

3 DISCUSSION

The resulting model will be the basis for the development of a semiautomatic tool, directly connected with the ZmapujTo application allowing for the processing of reports. Processing would include the investigation of the probability that a given dump can or cannot occur in a given area. For data control, some threshold would be set. In the case that the probability of occurrence is lower than the set value, then manual administration would be triggered. In the other case, a report would be automatically forwarded to an appropriate municipal or city department that is responsible for the agenda of illegal dumps. It is necessary to note that this tool would solve only spatial accuracy, meaning if it is possible that an illegal dump is occurring in the given place or not. Disadvantages of automatic forwarding occur when an attribute accuracy check would take place. These attributes involve situations when people report only a minor mess or overfilled dustbin as an illegal dump. At the same time, the existing concept faces the problem of accurate localisation of the dump.
Especially in dense urban areas and forested areas, accurate localisation using a smartphone is highly problematic. Location can be determined with a +- 10 meter inaccuracy as illegal dumps can be placed in the middle of a road, in a stream, or the centre of a building. The resulting model tries to eliminate spatial error using a focal circle filter with a 10m diameter.

Nevertheless, placement inside a building or house does not necessarily mean an error in the data. In the case that the mobile phone does not enable connection using data packets, the ZmapujTo application enables a person to record the report in the phone's memory and send it later using a Wi-Fi connection. This may cause that the location of submitting the report will be recorded. This error can be eliminated by manual localisation directly on the map, also enabled by the application. Computation of probability of occurrence of illegal dumps for big cities in the Czech Republic as well as for the entire Czech Republic would be highly problematic. Not all cities share the same attitude towards data as Prague. A high computational demand presents another issue. A stand-alone model of occurrence of illegal dumps can be useful for particular quarters of Prague; for the detection of places where a higher risk of occurrence of an illegal dump exists.

4 CONCLUSION

Crowdsourcing has rapidly developed due to an increasing number of people with access to the internet. Many people possess smartphones with a data connection to the internet, not only limited to Wi-Fi networks. This is one of the most important prerequisites for the development of mobile applications, which are a huge source of spatial data based on crowdsourcing principles. A system for the assessment of spatial accuracy is only one part of a complex tool for reporting quality assessment accepted by the ZmapujTo application. Another integral part is the concept of assessing attribute accuracy. Assignment of the correct category to the report is crucial (such as illegal dump, minor mess). In the case of illegal dumps, the application allows a person to enter the weight: Less than 10kg - small, less than 50kg - medium, large. This assessment is very subjective - weight estimation can be problematic for users. Due to the given reasons, a decision scheme for the correct classification will become part of the final application. In the case of applying the system to real situations, there is an expectation that 90% of large illegal dumps will be automatically assessed and handed over to authorities. This will save administrators time because, until now, all reports were assessed manually.

The Planning and Development Institute of Prague is also interested in the outputs. Maps can also serve for occurrence prediction. The Planning and Development Institute of Prague can then detect places where illegal dumps occur more frequently compared to other areas of Prague.

Apart from major applications such as OSM, which use quality assessment based on complex algorithms, this method is strictly based on solutions using the GIS approach. Using GIS for these types of tasks is not very common, therefore, the goal of this paper is to introduce possibilities how to successfully use GIS in the given field. This tool is a typical case of use of simpler analytical tools for the creation of complex tools for the assessment of spatial data quality. Successful connection of crowdsourcing and GIS opens new areas of geoinformatics. Geoinformatics possess significant potential for future utilisation in the given domain.

5 REFERENCES

Brown, L., Jason, 2015. SDMtoolbox. http://sdmtoolbox.org/ (last accessed 11.4.2016)

Brus, J., 2013. The role of standards in spatial data quality visualisation, International Multidisciplinary Scientific GeoConference Surveying Geology and Mining Ecology Management, SGEM, pp. 571-578.

Dilts, T., E, 2015. Prepare Rasters for Maxent Tool for ArcGIS 10.1. http://www.arcgis.com/home/item.html?id=11bf7e689c92413f8d31933b3e1f56b1 (last accessed 12.2.2016)

Exel, v., M, Dias, E., Fruijtier, S., 2010. The impact of crowdsourcing on spatial data quality indicators. Proceedings of GiScience 2011.

Heer, J., Bostock, M., 2010. Crowdsourcing graphical perception: using mechanical turk to assess visualization design, Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM, pp. 203-212.

Howe, J., 2008. Crowdsourcing: Why the power of the crowd is driving the future of business. Crown Business, New York.

Idris, N.H., Jackson, M., Ishak, M., 2014. A conceptual model of the automated credibility assessment of the volunteered geographic information, IOP Conference Series: Earth and Environmental Science. IOP Publishing, p. 012070.

Kubásek, M., 2015. Civic Issues Reporting and Involvement of Volunteers as a Phenomenon in the Czech Republic, Environmental Software Systems. Infrastructures, Services and Applications. Springer, pp. 151-159.

Kubásek, M., Hřebíček, J., 2013. Crowdsource Approach for Mapping of Illegal Dumps in the Czech Republic. International Journal of Spatial Data Infrastructures Research 8 144-157.

Merow, C., Smith, M.J., Silander, J.A., 2013. A practical guide to MaxEnt for modeling species' distributions: what it does, and why inputs and settings matter. Ecography 36(10) 1058-1069.

Phillips, S.J., Dudík, M., 2008. Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. Ecography 31(2) 161-175.

Rice, M.T., Paez, F.I., Mulhollen, A.P., Shore, B.M., Caldwell, D.R., 2012. Crowdsourced Geospatial Data: A report on the emerging phenomena of crowdsourced and user-generated geospatial data. DTIC Document.

Sarma, A., Parameswaran, A., Widom, J., 2014. Optimal Worker Quality and Answer Estimates in Crowd-Powered Filtering and Rating, Second AAAI Conference on Human Computation and Crowdsourcing.

Stark, H.-J., 2010. Quality assurance of crowdsourced Geocoded address-data within openaddresses. Salzburg University.

Sui, D., Goodchild, M., Elwood, S., 2013. Volunteered geographic information, the exaflood, and the growing digital divide, Crowdsourcing geographic knowledge. Springer, pp. 1-12.

Xia, J., 2012. Metrics to measure open geospatial data quality. J. Issues Sci. Technol. Librarianship.