Machine learning techniques for inspection data analysis

Rajko Bulatovic*, Zora Konjovic**, Aleksandar Ivic***, Djordje Obradovic***
*Provincial Secretariat for Health, Vojvodina, Serbia
** Singidunum University, Belgrade, Serbia
*** Faculty of Technical Sciences, University of Novi Sad, Serbia
rajko.bulatovic@vojvodina.gov.rs, zkonjovic@singidunum.ac.rs, ivicaca@gmail.com, obrad@uns.ac.rs,

***Abstract* - The unique data set was created by collecting data about sanitary inspection control on the territory of AP Vojvodina during the period of two years. The total number of multi-dimensional data records is 28,403. The Open Data with complete details is published on web site of Provincial Health Secretariat in the section Documents – Data, available for a bulk download. Data comprise temporal, spatial and categorical components and as such are highly suitable for a variety of analyses by means of machine learning techniques, especially neural networks. In this paper examples of linear regression and neural networks applications to analysis of the data are presented. The obtained results can be used for improving daily tasks like estimating inspection control workload, and alike.**

With the adoption of the Law on Inspection Control (in further text referred as Law) in 2015, Serbian Parliament opened the way for a paradigm shift in government inspection services. In fact, only with the adoption of this Law the preconditions for fundamental reform of governmental inspection services were established. The adoption of the Law was preceded by analysis, study visits to countries that cultivate a good inspection practice, and consulting local and foreign experts' consultation in the process of passing the Law. The result was the Law that was positively evaluated not only by non-governmental organizations but also by the European Commission.

One of the mechanisms which create the preconditions for the lawful and proper operation of the inspection managers and inspectors are proper information systems and software solutions that enable the transparent work of inspections, harmonization of inspection practices, timely access to data, case monitoring, risk evaluation, and corruption prevention. [1]

An important new aspect of the Law is regulation of the transparency of governmental inspection services. The Law defines, but does not restrict, activities aimed at establishing transparency, e.g. publication of inspection control plans and reports , inspection checklists, best and worse rated subjects of inspections (white and black lists), and current regulation concerning inspections on public internet sites. The practice of publishing required information is not widespread at the moment among inspections in Serbia, mostly due to the lack of technical capacity. This paper proposes publishing additional non-mandatory data extracted from the database and/or derived from the raw data. Furthermore, the conducted data analysis could be the base for resource planning in future inspection supervisions.

In 2011 the Provincial Government of Vojvodina enjoined the Provincial Department of Health to prepare the tender documents for the implementation of the software for sanitary control. The goal was a software system aimed at improvement of the entire inspection process. More precisely, in order to achieve this global goal, the software system should provide for:

- Automated creation of documents with electronic archive functionality.
- Efficient creation and maintenance of the following up to date electronic databases:
  - Registry of objects under sanitary control,
  - Registry of persons under sanitary control,
  - Registry of relevant regulations and legislation used in the control.
- Monitoring of sanitary and hygiene conditions in facilities under sanitary supervision.
- Significant improvement of the functions of planning inspection control, autonomous processing of reports, monitoring of inspection process (current outcome and legal assistance to the inspector from start to the end of the control process).
- Assisting in establishing a uniform inspection methodology and procedures.
- Equal treatment for all subjects under supervision.
- Creating the conditions for a comfortable, efficient, precise, legally safe operation of inspectors.
- Supervision of the work and expertise of inspectors.
- Internal licensing (periodic proficiency testing) for inspectors.
- Preventing illicit relations, corruption, and other adverse consequences.
- Raising transparency of inspection by publishing information regarding regulations, practices, lessons about appeals, answers to frequently asked questions, educational materials and the like.

Good international practice in the organization of the inspection that is sublimated in the recommendations of the World Bank presented the study: "Good Practices for Business Inspections - Guidelines for Reformers, World Bank Group / Small and Medium Enterprise Department – 2006" which was the starting basis for requirements specification. In 2012 the team from the Faculty of Technical Sciences in Novi Sad implemented the software for sanitary control with requested features. The software is in the production phase since January 2013 on the territory of AP Vojvodina.

The software has won awards in 2014 by the Informatics Society of Serbia in the field of innovative application of IT in the business environment.

Sanitary control process are implemented with a weekly work plan, which is designed using the module for planning. This module was enriched in 2015 with the addition of smart planning of inspection control based on dynamic risk level based on AI techniques. The system for smart planning is described in the paper "Intelligent planning of inspection control based on a dynamic level of risk" [2], an award winning paper of the Infofest conference 2015.

The methodology of inspection oversight, from the very beginning of the system use in 2012, relies upon checklists, which are also a guide to performing the inspection. Provincial Secretariat for Health, Social Policy, and Demography created more than 120 different types of checklists for the most common types of facilities that are subject to sanitary surveillance. Checklists are in line with current legislation. Answers to inquiries automatically calculate the level of compliance of the supervised entities with the law. Software tools enable the search and processing of all parameters that are used in the control lists queries.

The software is an essential IT tool for the work of sanitary inspectors. All users of the system, including around 80 sanitary inspectors, passed individual and group training.

Serbia started the implementation of an Open data Initiative in 2015. The Strategy for the Development of e-Government comprises the introduction of an open data policy. Therefore, as the first step towards that goal, with the support from the World Bank and UNDP, the Government of Serbia conducted an Open Data Readiness Assessment (ODRA) [3] in June-November 2015, upon the request of the Ministry of Public Administration and Local Self-Government, Directorate of e-Government. The assessment concludes that Serbia is in a good position to move forward with an Open Data Initiative, although stronger senior management level support and awareness are needed.

Strategy for Development of e-Government and ODRA recommendations anticipate the next step of the open data initiative in Serbia as the development of an open data portal that would provide, via a metadata catalog, a single point of access to data of the Government institutions, agencies, and bodies for anyone to reuse. At the moment of writing this article, the national portal is in final phase of development, available on web address http://data.gov.rs/sr/. So far, the portal hold the data from six governmental institutions that already published data sets as open data, including the Ministry of Education, Science, and Technological Development; Ministry of Interior; The Public Procurement Office; The Commissioner for Information of Public Importance and Personal Data Protection. This is the first data set regarding inspections data released in Serbia. Data is available for a bulk download on the official web site of Secretariat for Health of Autonomous Province of Vojvodina [4].

One successful example of making similar data (those for restaurant inspection) publicly available is an initiative from the last century. Since the mid-1990s, NYC has been posting its restaurant inspection data online to reach a wider audience [5]. Outcomes of food safety inspections of restaurants and other similar providers of food to the public in Open Data form are now widely accepted. On web portal of The US City Open Data Census [6], more than 70 open data sets regarding restaurant inspections are available. Relaying upon that data arise numerous and various applications, including a mobile app to help user chose a nearby place to eat. Another example is HDScores which gathers health inspection data from across the country and incorporates data from 530,000 establishments. Need to gather data from various system led to proposing a new standard for Local Inspector Value-Entry Specification (LIVES) [7].

I. DATA

Collecting data used in this study was done by checklists that are an integral part of the software. The initial checklists were swapped in 2016 with a set of checklists approved by the Ministry of Health of the Republic of Serbia and Coordination Committee of the Ministry of Public Administration and Local Self-Government. Checklists are available on the website of the Ministry of Health as well as on the website of the Secretariat of Health.

Checklists used by the Provincial Sanitary Inspection are additionally customized for electronic processing, while the contents and questions are fully aligned with official checklists. Thus decorated checklist as Excel workbooks contain several worksheets and enable collection of data on the supervised entity, its organizational units, responsible and present representative. A special worksheet contains scored questions for an inspection check, which allow automatic calculation of identified risk in business operations of supervised subjects during the supervision. Data are collected in the process of field supervision. Sanitary inspectors fill checklists during the assessment of the compliance and security risk.

The data used in this paper have been collected by 65 sanitary inspectors in the period January 2013 to December 2016 on the territory of the Province of Vojvodina. The electronic archive spanning the period 2013 – 2016 stores over 45 000 filled checklists and over 80 000 cases formed by inspectors, containing more than 500,000 documents that are stored and available in the software for sanitary control. The total number of records in a DataSet is 28,403 collected in the period 1.1.2015. - 01.01.2017.

Table 1 presents semantics and some quality indicators of collected data.

The field *DatumInspekcije* is the date when inspection took place. Fields *Zapoceta* and *Zvrsena* are times of inspection control beginning and ending respectively.

The field *Inspector* is the identification of inspector in charge. Inspectors have territorial jurisdiction, defined by the administrative boundaries of administrative districts in

the province. Each administrative district has department and representative of Sanitary Inspection.

Fields *Mesto*, *Opstina*, and *Okrug* describe the spatial characteristics of the data set.

The field *Kontrolna lista* is a document containing the checklist for inspection control.

The field *VrstaPregleda* indicates inspection control type.

The field *Usaglasenost* is a percentage of compliance in inspection control.

The field *Kategorizacija* is defined by the Ministry and represents the category of the epidemiological risk.

The field *VrstaMin* is a hierarchical categorization of business activity prescribed by the Serbian Ministry of Health.

The field *OJVrsta* is extended classification that accurately determines the object affiliation to a group of objects that are under sanitary control.

Table I.

DATA SET ONTOLOGY AND QUALITY INDICATORS

Name	**Type**	**Format**	**Missing value**	**Distinct value**
DatumInspekcije	Date	dd.MM.yyyy.	0	N/A
Zapoceta	String	HH:mm	1.627	644
Zavrsena	String	HH:mm	1.939	636
Inspektor	String		0	60
Mesto	String		773	608
Opstina	String		4.369	160
Okrug	String		4.374	228
Kontrolna lista	String		0	102
Usaglasenost	int		25.984	8
VrstaPregleda	String		2.411	13
Kategorizacija	String		5.946	6
VrstaMin	String		4.182	138
OJVrsta	String		4.426	298

Even simple processing of the data provides information for getting useful insight in the inspection process and actors. They are presented by Table II and Figures 1 and 2.

TABLE II
INSPECTION START/END TIME DISTRIBUTION (TOP TEN VALUES)

Position	Start time		End time	
	Value	Frequency	Value	Frequency
1	10:00	2295	null	1939
2	11:00	1665	12:00	1654
3	null[1]	1627	11:00	1425
4	09:00	1505	13:00	1370
5	08:00	1467	12:30	1124
6	12:00	1267	11:30	1064
7	10:30	1150	14:00	1051
8	09:30	902	10:00	947
9	11:30	826	10:30	882
10	13:00	706	13:30	785

Figure 1 represents the date distribution of inspection control for a whole period.

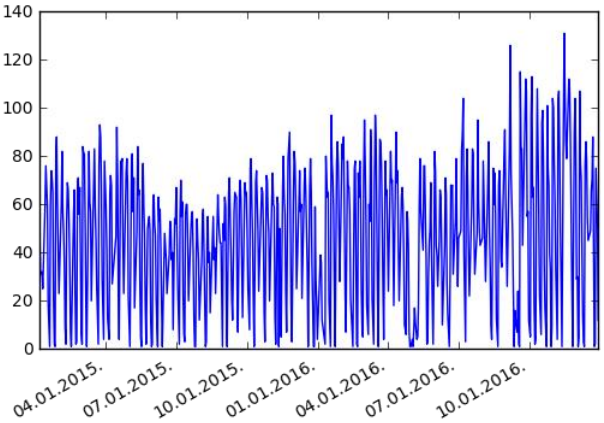

Figure 1 – Date distribution of inspection controls

Depersonalized code for inspector in charge can be used to characterize a behavior of inspectors.

The graph in Figure 2 shows the sum number of inspections per inspector.

[1] null value appears if inspector omits to fill the field

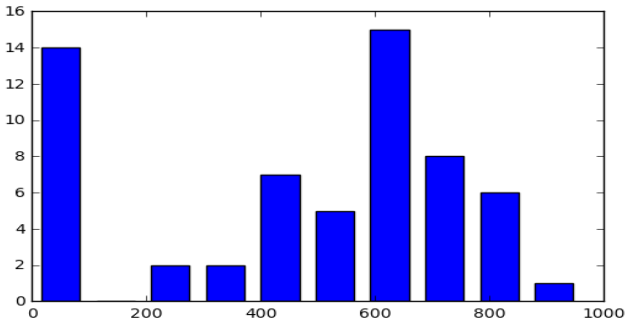

Figure 2 – Number of inspections per Inspector histogram

The chart shows that by criteria of number of inspections there are two types of inspectors: inspectors who have a lot of control and a few inspectors with a small number of control (several months of absence from work due to sick leave, complex forms of control carried out by inspectors' advisers or heads of departments). The average value of about 470 supervisions may mislead. The average value of each group shows that average in the first is about 50 supervisions while another group average is about 600 inspections. Grouping these data could be made by using the k-means algorithm with the number of Group 2.

The number of inspections is not the only measure of inspector engagement. More objective measurement is to calculate the sum of all controls duration. With **control duration** as criteria, the engagement of inspectors should be relatively uniform, because the inspections with a smaller number of controls usually have inspections that lasts longer. The duration of contol is calculated as a difference of Inspection end and start time. There is no information in the dataset to indicate the situation where inspection lasts more than one day.

As mentioned, fields *City* (*Mesto*), *Municipality* (*Opstina*), and *District* (*Okrug*) describe the spatial characteristics of the data set. Two origins are used for this data: the official registry of the Agency for Business Registers and manual entry of the inspectors. Hence, in these data mistakes are expected to occur like: "Novo Bečej" instead of "Novi Bečej"; "Krušedolska Prnjavor" (non-existent place); "R. Krstur" is the same as "Ruski Krstur"; "Novi Sad" (with the double space between).

The analysis of these data could show all types of errors, and the machine learning algorithms could be trained to find frequent mistakes. Training data set for the AI algorithms could be the registers of the names of places, municipalities, and districts in the province of Vojvodina. Trained algorithms could be applied for the correction of abovementioned names.

The Naïve Bayes Filter is appropriate for prediction and planning of activities and number of controls. Regression algorithm could be used for correlation between the numbers of inhabitants with the number of controls in the area.

Field *Control list* **[**Kontrolna lista**],** which is a document containing the checklist for inspection control, for the dataset described in the paper consist of two control list types. The first type is marked with Q_00.01.XX - 09.01.XX and the other with Q_10.01 - 10.31. The second one, made by Ministry of Health of the Republic of Serbia, is applied since 30th of April 2016.

Field *Type of inspection control* (*VrstaPregleda*) should take one of the following values: **ordinary**, **extraordinary**, **control**, **supplementary** However, analysis of the data revealed that there are anomalies, like in abovementioned field *City*.

Very accurate prediction of inspection duration could be achieved with the usage of this data together with the data from *Control List* and *Type of Object*.

Field *Categorization* (*Kategorizacija*), which is defined by the Ministry and represents the category of the epidemiological risk, is a basis for planning the frequency of inspection within a period (calendar year). The analysis of the data set shows that majority of the inspections apply to objects from Category III, which represent 74% of buildings under sanitary control. A smaller number of the checks are in Category II objects, which represents 20% of buildings, and at the end in Category I, which makes 6% of buildings under control. Further analysis could be done to connect the checklist, the duration of inspection and determining the category. A neural network trained by lists and length of control could be used to determine the category. In addition to neural networks algorithms, K-nearest neighbor and naïve Bayes filter is the option.

Type of building (*OJVrsta*) is an extended classification that accurately determines the object affiliation to a group of objects that are under sanitary control, while the field *VrstaMin* determines business activity as one of three major groups: A health care; B manufacturing and marketing; C water supply. These two fields together could be used to train machine learning algorithms aimed at prediction of co-occurrence of a particular style *VrstaObjekta* and *VrstaMin*.

II. EXAMPLES OF DATA UTILIZATION

In this section we present three examples of application of machine learning techniques to dataset described in previous section supplemented with open data available in Serbia.

The first two examples illustrate a possibility of inspection controls planning based on density of population, while the third one is about prediction of inspection control duration.

A. First example: A simple regression model

In the first example, the correlation between a number of residents in the particular area and the number of inspections is determined. This correlation is useful for planning inspection controls. The data originate from two datasets. In addition to the dataset described above, the data on municipalities/census [8] were used. Since Statistics published data as Excel workbooks, the additional effort has been put to put it in machine-readable format. As a foreign key for connecting datasets, the Name of the municipality in Latin letters is used. Two methods of machine learning are used for that purpose, regression model and artificial neural network.

The correlation between number of inspection controls and residents' number is shown on Figures 3 and 4.

The graph is produced with Python and NumPy package for scientific computing.

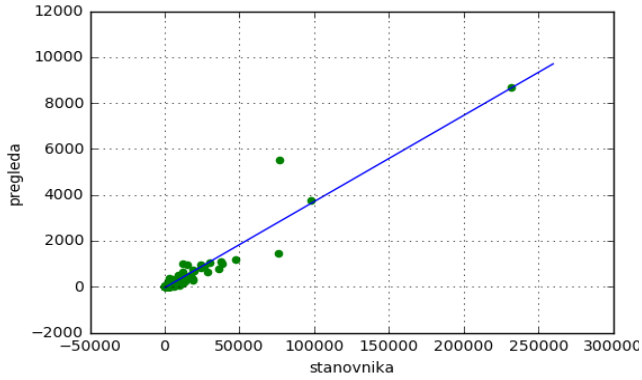

Figure 3 – Correlation of inspection controls with resident number

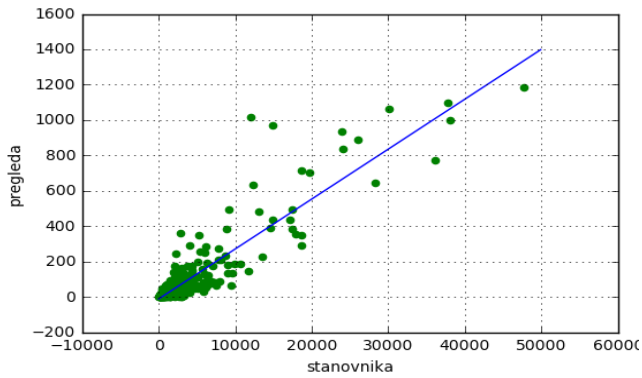

Figure 4 - Correlation of inspection controls with resident number zoomed

The result of machine learning are the parameters of the linear regression which are the basis to determine a linear relationship between the number of inhabitants and the number of completed inspections. Linear regression is one of the simplest algorithms for approximation.

Imprecision in the input data led to interesting higher order polynomial approximation. Figure 6 shows an example with six-degree polynomial approximations. The data fits surprisingly well in an approximation function. However, it can be seen that due to the complexity of the 6th degree polynomial the generalizations is lost which led to an incorrect result.

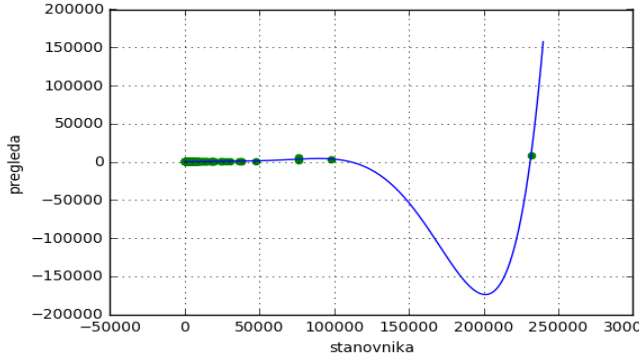

Figure 5 – 6th degree polynomial model of correlation of inspection controls with resident number

B. *Second example: Regression modelled by artificial neural network*

The same problem was solved applying Artificial Neural Network. The network is organized as a multi-layer perceptron with following characteristics:
Input layer: 1 neuron
Hidden layer 1: 5 neurons, Activation function TanH
Hidden layer 2: 5 neurons, Activation function TanH
Output layer: 1 neuron, Activation function ReLU
Training function:
SGD (Stohastic Gradient Descent),
learning rate: 0.2,
decay: 0.000001,
momentum:0.7,
loss function: mean_squared_error

Figure 6 shows network training error, while Figure 7 shows correlation gained through ANN.

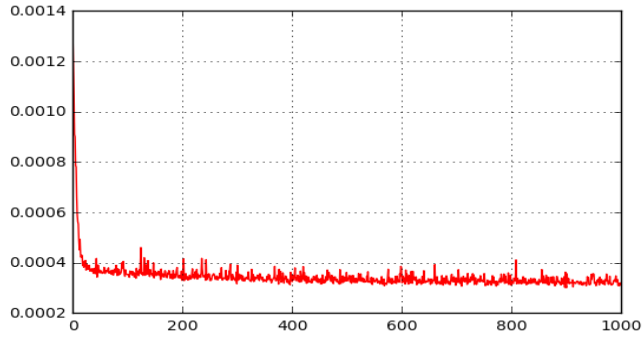

Figure 6 – Training error

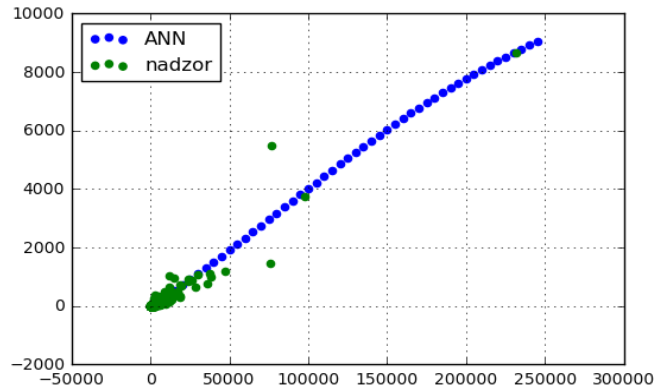

Figure 7 – Neural network for modelling correlation of inspection controls with resident number

C. *Third example: Prediction of Inspection control duration*

This example presents a neural network applied for prediction of inspection duration based on object category.

The original values of the field that represent object type [VrstaObjekta] are alphanumeric, so it is necessary to perform a conversion to the form adequate for the training of the network. For that purpose objects are transformed to a vector with the length equal to a number of possible values of the object (n). All coordinates of this vector are 0, except the one that corresponds to ordinal number *VrstaObjekta*, and that is 1. Output values are also

transformed in analogous way: mean subtraction Y -= Y.mean(), and then normalization Y /=Y.std().

Network is organized as a multi-layer perceptron with following characteristics:

Input layer: 1 neuron
Hidden layer: 50 neurons, Activation function TanH
Output layer: 1 neuron, Activation function TanH
Training function:
SGD (Stochastic Gradient Descent),
learning rate: 0.1,
decay: 0.000001,
momentum: 0.7,
loss function: mean_squared_error

After performed training the error change is displayed on the Figure 8.

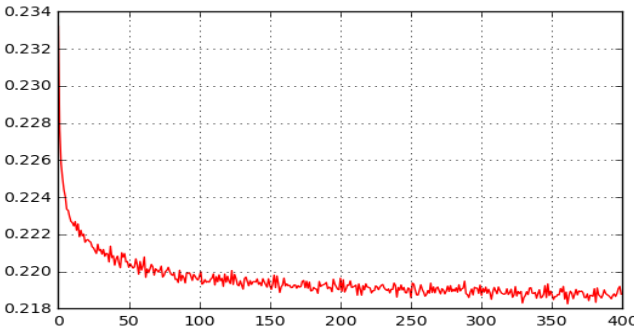

Figure 8 - Training error

The results for the most important object categories are displayed in Table II.

TABLE II.

RESULT OF THE NEURAL NETWORK TRAINING

Object type	**Inspection control duration (min)**
B301 Klinika	134
B211 Opšta bolnica	119
E101 Hoteli i moteli sa restoranom	117
F101 Osnovno obrazovanje opšte	109
E222 Kuhinje kolektivne ishrane	106
B101 Dom zdravlja	102
E204 Picerija	98
D101 Prodavnica	81

III. CONCLUSION

This paper presents the data set containing the data collected during two years of sanitary inspection control on the territory of Province of Vojvodina. The data available in Open Data format were extracted from the database established in 2012 relying upon the flexible information system that supports business process and legislation changes, and smooth implementation. Some deliberate approach to information system design and implementation are that the inspector can type in some data without restrictions/control. Fortunately, this did not produce critical consequences in core business processes functionality; on the contrary, they prove to be very useful for reporting and controlling purposes.

The paper also argues about advantages of the open data approach and presents illustrative examples of AI and machine learning applications utilizing such data. The quality and amount of the data are essential for utilization of efficient data mining methods/algorithms, while proper metadata and the processes of data collection are required for the analyst to understand the data better. In our example the quality of data is high, which is warranted by the data provider/owner, while its utilization should be warranted by data openness and content which is interesting for citizens, number of business organizations in different industries, and numerous administrative entities.

The examples provided in this paper are only illustrations of applications that can be implemented by applying selected data analysis algorithms to the inspection data set as well as to data obtained by its integration with other open data. New applications are the matter of available open data and users' creativity.

The data set presented in this paper allows for defining a procedure aimed at quality assurance of algorithms for lexical analysis and correction of textual input data like manual filling of the address field, identity management algorithms, and dynamic risk assessment based on previous experience and data. These are, by authors' opinion, some up-and-coming directions for the further research in the field addressed by this paper.

REFERENCES

[1] M. Stefanović, D. Milovanović, J. Stefanović, and I. Dragošan, "GUIDE for the application of the Law on Inspection Control," Inspection supervision - Serbia - Legislation ISBN 978-86-919327-0-1, Nov. 2015.

[2] R. Bulatović and Đ. Obradović, "Inteligentno planiranje inspekcijskog nadzora bazirano na dinamičkom stepenu rizika," presented at the INFOFEST, Miločer, Montenegro.

[3] A. Zijlstra, O. Petrov, and A. Ivic, "Open Data Readiness Assesment," UNDP, WB, MoPALSG, DEU, Republic of Serbia, Dec. 2015.

[4] "Sanitary Inspection | Open Data download." [Online]. Available: http://www.zdravstvo.vojvodina.gov.rs/index.php?option=com_docman&task=cat_view&gid=121&Itemid=91. [Accessed: 14-Apr-2017].

[5] N. Helbig, A. M. Cresswell, G. B. Burke, and L. Luna-Reyes, "The dynamics of opening government data, CS - Restaurant health inspection data in New York City (NYC)," *Cent. Technol. Gov. Available Httpwww Ctg Albany Edupublicationsreportsopendata*, 2012.

[6] "Restaurant Inspections — Datasets - US City Open Data Census." [Online]. Available: http://us-city.census.okfn.org/dataset/food-safety. [Accessed: 30-Jan-2017].

[7] "Local Inspector Value-Entry Specification - Yelp." [Online]. Available: https://www.yelp.com/healthscores. [Accessed: 30-Jan-2017].

[8] "Download Serbia cesus results 2011." [Online]. Available: http://popis2011.stat.rs/?page_id=2162&lang=en. [Accessed: 14-Apr-2017].