ON THE ANALYSIS OF DISCRETE DATA – FINDING DEPENDENCIES IN SMALL SAMPLE SIZES

S. Jozova[*], *M. Matowicki*[*], *O. Pribyl*[*], *M. Zachova*[*], *S. Opasanon*[†], *R. Ziolkowski*[‡]

Abstract: An analysis of survey data is a fundamental part of research concerning various aspects of human behavior. Such survey data are often discrete, and the size of the collected sample is regularly insufficient for the most potent modelling tools such as logistic regression, clustering, and other data mining techniques. In this paper, we take a closer look at the results of the stated preference survey analyzing how inhabitants of cities in Thailand, Poland, and Czechia understand and perceive "smartness" of a city. An international survey was conducted, where respondents were asked 15 questions. Since the most common data modelling tools failed to provide a useful insight into the relationship between variables, so-called lambda coefficient was used and its usefulness for such challenging data was verified. It uses the principle of conditional probability and proves to be truly useful even in data sets with relatively small sample size.

Key words: *lambda coefficient, discrete data analysis, soft computing*

Received: September 11, 2020
Revised and accepted: October 30, 2021

DOI: 10.14311/NNW.2021.31.017

1. Introduction

In all kinds of behavioral studies, preference and attitudinal surveys are typically the fundamental data collection methods. While it introduces many advantages, such as low cost of study and easy survey preparation and dissemination, this method bears also several drawbacks, such as low reliability of sample collection, mainly discrete data collection, and very often low response rate of a complete survey [1]. As a result, the collected data often consist of a considerable number of discrete variables and fairly small sample size. This combination makes the application of standard mathematical tools used in data modelling and prediction a very complicated task [2,3]. In this study, we present such an example of survey data of inadequate size for common data modelling tools. Given that the performed survey provided only binary and discrete values, available, sample size proved to be

[*] Sarka Jozova, Michal Matowicki – Corresponding author, Ondrej Pribyl, Michaela Zachova; Faculty of Transportation Sciences, Czech Technical University in Prague, Czechia, Na Florenci 25, Prague, Czechia, E-mail: matowmic@fd.cvut.cz

[†] Sathaporn Opasanon; Thammasat Business School, Thammasat University, Thailand

[‡] Robert Ziolkowski; Department of Civil and Environmental Engineering, Bialystok University of Technology, Poland

insufficient for neither logistic regression, clustering nor decision trees [4]. Dolnicar et al. published a study in 2014 recommending a sample size between $60k$ and $70k$, where k equals the number of observations (different values of dependent variable). Dolnicar references an even older study suggesting a sample size guideline of 2^k observations, where k equals the number of features [5]. This particular technique also seems to be a potential strategy to manage feature "creep" – rapid increase in number of features examined in a model. If 2^k observations far surpass the observation count, the analyst may also consider reducing the initial feature set in lieu of increasing the sample size. These modelling methods were unable to identify underlying data relationships with a statistical significance level of $p = 0.05$[1] or lower in our data set. This led to the authors investigation of robust analysis methods applicable for identifying dependencies between variables in datasets. According to our careful literature analysis, a coefficient is used based on the theory of conditional probability. The purpose of this paper is the demonstration of new proposed analysis method for discrete data, on example of the data collected in international survey on attitudes and believed definitions of a Smart City.

This paper has the following structure. Following the introduction, a literature review and state-of-the-art analysis are provided. In Section 3, the performed data collection and research topics are introduced. Section 4 describes the applied methodology of data analysis and the proposed lambda coefficient. The final section provides the results analysis, discussion, and conclusion of the study findings, respectively.

2. Analysis of discrete data

Our study resolves around the topic of the analysis of discrete data. The most common source of this data are questionnaires, which are commonly used in practice. Unfortunately, methods for the analysis of discrete data are not researched as well as for continuous data. Therefore, it is important to carefully and properly choose appropriate methods. Discrete data are analyzed with descriptive statistics [6] but we are particularly interested in finding relations between variables of discrete data. We want to determine who and under what circumstances chooses a particular definition of Smart City (SC), e.g., a Czech with a secondary education will choose the 4th definition with high probability.

Such relations can be searched using the Bayesian approach [7]. One of the procedures is as follows: one selected variable is used to create clusters, and then it is modelled inside these clusters. However, this method requires a sufficient amount of data, which are not available in our case.

To find the relations between individual variables x and y, it is possible to use one of the classification methods which operates with discrete data. In general, these methods make it possible to determine the output of y according to the classification of variables x at any given time t.

One of the classification methods for discrete data is a neural network [8], which allows any relation to be well approximated by a network with a certain

[1] The data set was experimentally validated on the selected methods with the chosen significance level

complexity and a sufficient amount of training data. Therefore, it is a suitable method for comparing the results obtained from other methods. However, the necessary condition is a sufficient amount of data. The disadvantage is that the structure of the created network cannot be well decoded and thus determination of the significance of individual variables for classification is (to a large extend) unknown. The quality of the neural network model can be verified by the accuracy of data classification [9].

Decision trees [10] work in a similar fashion. They have the ability to describe even complex relations, but they tend to converge to a very complex tree structure. Another method worth mentioning is logistic regression [11]. It is primarily used for classification (prediction) of a discrete variable. Nevertheless, if the discrete data is converted to dummy variables, it is possible to effectively use regression for this data as well [12]. All three methods are suitable for the classification of discrete data, but not always for their analysis and finding relations between variables. The overview of their strengths and weaknesses is depicted in Tab. I.

Model	Decision Tree	Neural Network	Logistic Regression
Application scope	Classification and prediction	General purposes such as classification, estimation, prediction, segmentation, association and others	Discrete choice prediction
Estimation method	Recursive partitioning	Backpropagation	Maximum likelihood
Estimation time	Fast	Extremely slow	Moderate
Interpretability	Explicit decision trees / *if-then* rules	Implicit “black box”	Explicit utility functions

Tab. I *Overview of the capabilities of chosen data mining techniques [13].*

For this purpose, the lambda coefficient *lambda* or Goodman–Kruskal lambda is used [14]. This method tests the relation of individual variables x on y, i.e. the set $x_i \to y$. The principle of the test is to compare the quality of the prediction y with and without knowledge of x, i.e. from empirical probability functions $f(y)$ and $f(y|x)$. Where lambda is practically zero, the dependence does not exist or is completely negligible. The lambda coefficient λ was successfully used in the article [15] in the field of data analysis using Bayesian networks and conditional probability functions. The dependencies found between the variables were checked by using this method. Furthermore, the classification of community data into original and new clusters was compared in the literature [16] also by using Goodman–Kruskal lambda.

3. Performed study

Data for comparative analysis of quality of living in smart cities were obtained from an online questionnaire survey using the SurveyMonkey platform. Analysis of quality of living in smart cities was conducted in 2019–2020 for the Czechia, now broadened to Poland and Thailand with the purpose to determine differences and similarities between countries and generations.

The survey was translated into English and consisted in total of 15 questions, of which nine questions where aimed at identifying attitudes towards smart cities and quality of living, and the remaining six questions examined socio-demographic information on respondents. Each country shared the survey link.

In Czechia, the collection took place from November 2019 to January 2020. The survey collection link was disseminated by email to CTU students. It was published on social network profiles of 13 regional cities and published on Facebook profiles of two influencers reaching more than 3500 followers. In total, 273 responses were collected within 91 Czech cities.

In Poland, the collection took place in June and July 2020. The survey collection link was distributed by email to BTU students, universities, and municipal employees, it was published on social network profiles of 3 regional cities and published on Facebook. In total, 136 responses were collected within Polish cities.

In Thailand, with the power of social network, the questionnaire link was disseminated only two days (June 15, 2020) prior to the deadline (June 17, 2020) solely via LINE, a popular freeware messaging app on electronic devices. One of the difficulties in the data collection process lies in the fact that Thai people are non-English speakers. Hence, the respondents are limited to only Thai people who can understand English adequately well. Within two days of on-line data collection, 288 responses from Thailand were collected. It is important to note that all respondents are based in Bangkok and affiliated with Thammasat University as either lecturers, staff, current students, or alumni.

The questionnaire was distributed in these three countries, but the correspondents did not necessarily have to be nationals of those countries. Therefore, the fourth category of other countries of origin was created.

Fig. 1 shows the percentage of correspondents in relation to their country. The picture shows that Thailand has the largest representation, followed by Czechia, Poland, and a few correspondents from other countries, namely France, Germany or the UK.

The analyzed hypothesis of the study is that the believed aspects of the Smart City and preferred solutions applicable for Smart City influence the definition of Smart City by its inhabitants. In order to study this relationship, appropriate methods of discrete data analysis are applied as described hereunder.

4. Data and methods used

The data source is a questionnaire with 15 questions, which is described in the Appendix. This analysis includes only the part of the questionnaire that consists of 5 socio-demographic questions (Q1-*country of origin*, Q12-*gender*, Q13-*age*, Q14-*employment type*, Q15-*education level*) and 3 additional questions from the

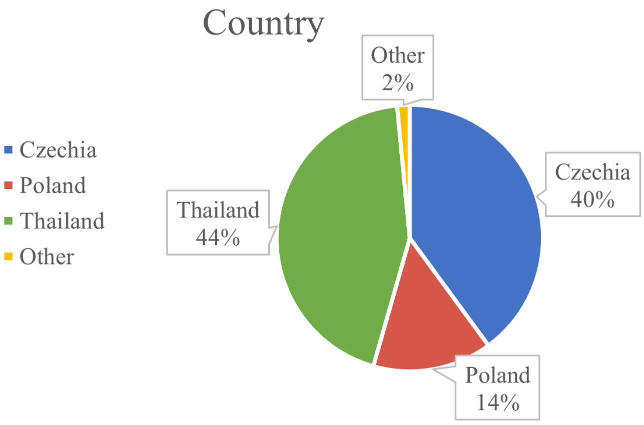

Fig. 1 *Correspondents – country of origin.*

category individual perception of Smart City (Q3-*definition of SC*, Q4-*aspects* and Q6-*solutions*). For our study, the collected data were preprocessed so that for each choice question values from 1 to n were assigned for each answer (n is the number of available answers for each question).

All data have either logical (binary) or discrete values, and thus the application of appropriate discrete analysis methods is needed. In our case, these (methods) are based on probability functions of variables, which are given as normalized histograms.

For model construction, various approaches were tested – i.e., data occupancy in individual model categories (output, regression vector, clustering variables and their subsets).

In our study, we assumed there is a relation between respondents' *definition of Smart City* (y) and socio-demographic characteristics of respondents, believed *aspects* and *solutions* (x) are sought after in Smart City. This analysis was performed on data previously filtered into homogenized clusters. In order to proceed with data analysis, it is necessary to determine whether a correlation between x and y exists. As the analyzed data are discrete, an appropriate analysis method must be used. An example of such a method is a neural network.

In the following analysis, we observe two links among the explanatory variables: (i) quality of the model created on the basis of the neural network, (ii) the strength of a link between y and x with help of so-called λ coefficient [17]. The purpose of data analysis is to determine the choice of definition.

If we take a closer look at the chosen definitions in individual countries, we can notice differences. Fig. 2 shows that the Czechs chose the definition *"Efficient city"* as the most common definition and the Thais voted dominantly for the fourth definition *"City for people"* as the best Smart City description, respectively. On the other hand, Poles dominantly perceived Smart City as a *"Technological city"* and *"Efficient city"*.

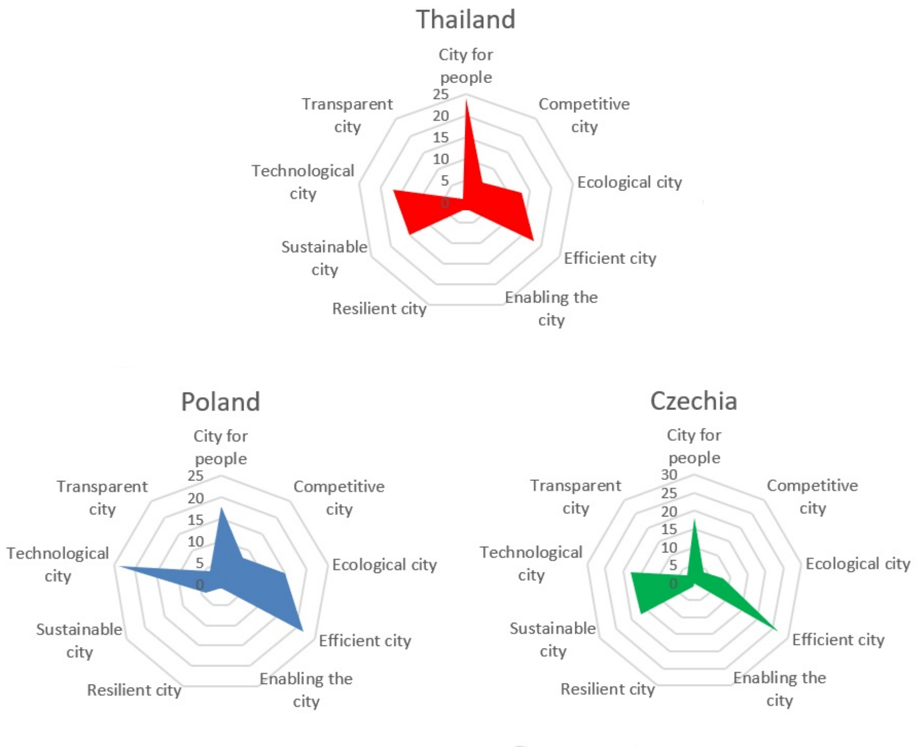

Fig. 2 *Choice of definition for individual states (each circular line corresponds to 5 % of respondents).*

4.1 Neural networks

Artificial neural networks (ANNs) are inspired by the biological nervous system to model the learning behavior of the human brain. They belong to the methods useful for approximating data patterns with a certain complexity and a sufficient amount of training data. In this paper, we focus on the so-called multilayer feedforward networks (MLFN), which is probably the most popular and widely used network type in many applications including forecasting. A basic scheme of a feedforward network is presented in Fig. 3.

Feedforward networks typically comprise three types of layers with neurons [18]: an input layer (the number of neurons in this layer corresponds to the number of inputs in the model), one or more hidden layers adding complexity to the possible mapping function and an output layer (containing nodes for each output variable). All neurons between the adjacent layers are interconnected and assigned weights. The hidden and output layer neurons, j, process their inputs x_i by multiplying each input by a corresponding weight coming from neuron i, w_{ji}, summing the product for all links to a neuron from the previous layer, k [19], and adding a bias

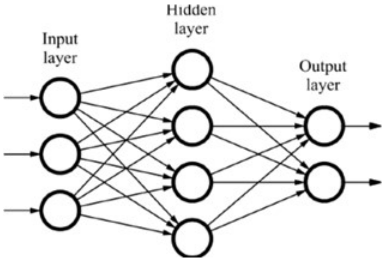

Fig. 3 *Example of feedforward single layer neural network.*

term z_j into a net input v_j:

$$v_j = \sum_{i=1}^{k} w_{ji} x_i + z_j. \tag{1}$$

The output of each neuron y_j is computed using a nonlinear transfer function θ (for example sigmoid function) on the net input v_j:

$$y_j = \theta\left(v_j\right). \tag{2}$$

Training of the feedforward network is done iteratively, using a backpropagation learning algorithm as gradient descent based on sum-squared error [20]. It adjusts particular weights between neurons using partial derivations in response to the error, E, between the output values of the network y_l and the expected output values a_l for all N training samples

$$E = \frac{1}{2} \sum_{l=n+1}^{N} \left(y_l - a_l\right)^2. \tag{3}$$

A nice overview of various applications of neural networks is presented for example, in [21, 22]. The authors clearly demonstrate the suitability of neural networks for classification of data that allows determining the existence of a relationship between quantities. This method, however, introduces a significant disadvantage. That is, the structure of the analyzed network cannot be well decoded and determination of the significance (dependence) of the influence of particular variables on the output behavior is virtually impossible.

With the classification of a model created with neural networks on filtered data, we investigate how the classification model works in each cluster. The quality of the neural network model can be verified by the accuracy of data classification [23]. This is expressed using the coefficient of accuracy, A, which is the number of correct classifications, N_C relative to the number of all samples, N:

$$A = \frac{N_C}{N}. \tag{4}$$

This parameter describes the accuracy of the model classification, which is exactly what we are interested in, in practice. If the accuracy is high, we know that a variable x affect variable y, but we are not able to assess which quantities of information do and which do not.

4.2 Conditional probability and Lambda coefficient

Conditional probability is the probability of an event A outcome being true given that an event B outcome is true, and is the key concept in Bayes' theorem. This is distinct from joint probability, which is the probability that both things are true without knowing that one of them must be true.

For example, one joint probability is "the probability that a respondent describes Smart City by definition 1 and at the same time his education level is Highschool" whereas a conditional probability is "the probability that a respondent describes Smart City by definition 1, given that his education level is Highschool" since adding information alters probability. This can be high or low depending on how frequently respondents with education level "Highschool" describe Smart City according to *definition 1.* For discrete random variables, the conditional probability mass function of Y given the occurrence of the value x of X can be written according to its definition as:

$$P_{Y|X}(y|x) = \frac{P_{Y,X}(y,x)}{P_X(x)}. \tag{5}$$

Dividing by $P_X(x)$ rescales the joint probability to the conditional probability. Since $P_X(x)$ is in the denominator, this is defined only for non-zero (hence strictly positive) $P_X(x)$. Furthermore, since $P_X(x) \leq 1$, it must be true that $P_{Y,X}(y,x) > 0$, and that they are only equal in the case where $P_X(x) = 1$. In any other case, it is more likely that $X = x$ and $Y = y$ if it is already known that $X = x$ than if that is not known. This concept is further developed and applied in lambda coefficient.

For the reasons mentioned above, we use the lambda coefficient λ, which tests the relation (dependency) of individual quantities x on y. The examined dependency is verified by the value of λ. In order to define this measure, let x and y be two nominal categorical variables with joint sample probabilities (proportions) P_{ij} and marginal probabilities $P_{i+} = \sum_{j=1}^{J} P_{ij}$ and $P_{+j} = \sum_{i=1}^{I}$ for $i = 1, \dots, I$ and $j = 1, \dots, J$. When represented in terms of an $I \times J$ contingency table with row variable x and column variable y, P_{ij} becomes the cell entry in row i and column j [24]. With x being explanatory (independent) variable and y the response (dependent) variable, the estimated lambda may be defined as

$$\hat{\lambda}_{x|y} = \frac{\sum_{i=1}^{I} P_{im} - P_{+m}}{1 - P_{+m}}, \tag{6}$$

where

$$P_{+m} = \max_j\{P_+j\}; \quad P_{im} = \max_j\{P_ij\}; \quad i = 1, \dots, I. \tag{7}$$

If lambda is close to or equal to zero, the dependence is negligible.

In order to improve the classification capabilities of our model, we applied homogenization of the analyzed dataset by filtering it with one of the value (E.g., we filtered the dataset by *Aspect 1* variable into 6 clusters – respondents who answered 1, 2, 3, 4, 5 or 6 respectively, see question Q4). For each data set of the filtered variable, we construct models:

$$f_c(y|x), \tag{8}$$

where c denotes individual clusters according to our filtering variable (in depicted example *education level*), y is the modelled variable (*definition of SC*) and x are explanatory variables.

In each cluster, the lambda coefficient (λ) asses how well each explanatory variable x explains the variance in y. Lambda has value $0 \leq \lambda \leq 1$ and follows as:

$$\lambda = (E_y - E_{xy})/E_y, \tag{9}$$

where E_y is the number of incorrect predictions in case only the $f(y)$ model is used for prediction, and E_{xy} is the number of incorrect predictions with the use of the $f(y|x)$ model (that is, when we use for prediction the information from the realized variable x).

With the help of lambda coefficient, we are able to determine the difference between the unconditional and conditional probability of the function $f(x)$. Therefore, the comparative prediction of y, without and with the use of the tested variable x respectively, is achieved.

4.3 Model construction

Model construction consists of the following steps:

1. A variable according to which we will filter our dataset is chosen;

2. Data are filtered into homogenized clusters by chosen variables: *gender*, *age*, *employment type*, *education level* and *country*;

3. The ANN is applied to verify whether filtered data hold a predictive potential (by calculation of accuracy);

4. The variable *definition of SC* is modelled (by λ coefficient) with the help of other Smart City *aspects* and preferred *solutions* of Smart City;

5. A comparison of conditional and unconditional probabilities is calculated to identify variables with impact on resulting $P(y)$.

In the first step, we identify a potential filtering variable that might bear feasible information for y prediction. With the help of this variable, we then perform filtering of the whole dataset based on the values of the chosen variable. This leads to the homogenization of the structure of the predicted data. As shown in Section 5, this improves the prediction capabilities significantly.

Step 3 is done iteratively for all selected variables according to which we wish to filter the raw dataset. In the last step, we compare the results of conditional probabilities in y prediction with the results of unconditional probability. By doing this, we can examine whether the additional information provided by the conditional probability of known variable x improves the predictive capabilities of our model.

5. The results

According to the presented methodology, in the first step a prediction with 3-layer feedforward neural network was performed. This analysis revealed that the believed *aspects* of Smart Cities, above all others, provide useful information for predicting the *definition of Smart City* among respondents. Moderately high prediction accuracy of $\mathbf{A = 0.681}$ expresses the relationship between *SC Aspect* variables group (*Aspect 1* through *Aspect 6)* and *Definitions of SC.* In order to improve the accuracy of the model, the analysis was performed also on data homogenized by socio-demographic information (i.e., *Education level*). The results in Tab. II prove that the accuracy was indeed improved.

Education level	1	2	3	4	5
Occupancy	143	87	138	218	56
Accuracy (A) [%]	88.8	94.2	92.0	82.1	100

Tab. II *Accuracy for the education cluster.*

Weighted average of the accuracy for the *Education* cluster equals to A = 0.889. Similarly, Tab. III was provided for all filtering variables, the achieved accuracy was transformed into the percentage of successful prediction.

Clusters	Education	Employment	Age	Gender	Country
Accuracy (A) [%]	88.9	85.2	85.1	80.5	84.1

Tab. III *Accuracy for all clustering variables.*

Achieved prediction accuracy (A) in filtered clusters reached over 80 %, which is a significant improvement from the initial 68.1 % in the unhomogenized dataset. Thus, it was decided that investigation of the relationships between individual SC aspects, socio-demographic characteristics, and predicted definition is advantageous in our study. However, although the high accuracy percentage confirms the existence of the information value in data, the neural network analysis does not provide us with tools to identify the localization or magnitude of the relationships between variables in our dataset. This was the motivation for the introduction of the lambda coefficient as a convenient parameter to investigate this.

In Tab. IV, we present the results of the lambda coefficient calculations for each aspect in individual clusters homogenized by the education parameter value for each definition. Furthermore, with bold text and shading of cells we highlighted values higher or equal than $\lambda = 0.1$. Such a value indicates that the examined combination of the filtered cluster and aspect variables bear information helpful in SC definition prediction. Values of the lambda coefficient which did not overcome 0.1[2], are irrelevant as the underlying relationships in the data are not adequately significant.

[2] Value chosen experimentally based on previous works

Aspects of SC	High School	Bachelor's Degree	Master's Degree	Ph.D.	Trade School
1. Transportation	0.082	0.081	0.058	0.079	**0.100**
2. Society	0.091	0.081	**0.106**	0.091	0.025
3. Environment	0.091	0.065	0.058	0.036	**0.125**
4. Economy	**0.127**	0.081	0.058	0.036	**0.100**
5. Buildings	0.045	**0.113**	0.087	0.055	**0.100**
6. Government	0.055	0.032	0.048	0.061	**0.125**

Tab. IV *Lambda coefficient values for the education cluster.*

To be able to determine which aspect influences the choice of the definition, we present the number representation of each definition answer in different clusters (Tab. V).

Definition of SC	1	2	3	4	5	6	7	8	9
Aspect 4 = 1	2	2	**6**	0	1	1	1	0	0
Aspect 4 = 2	1	4	7	**9**	5	0	0	1	0
Aspect 4 = 3	6	**14**	1	2	7	1	1	1	0
Aspect 4 = 4	2	4	**6**	4	**6**	1	0	0	3
Aspect 4 = 5	1	**7**	5	6	1	1	0	3	1
Aspect 4 = 6	0	2	4	**5**	1	1	0	0	0

Tab. V *Frequencies for the High School cluster and the Economy aspect.*

In Tab. V, the columns list all the choices for the *definition of SC* $(1-9)$, the rows list the selected *Aspect 4* (states that Smart City is economically efficient city) and its rating (all 6 aspects had to be ranked from least to most important in the questionnaire, so each aspect has a possible value $1-6$). Bold text in cells highlight the most populated answers for *Aspect 4*. Hence, Tab. V informs that respondent with high school education level, which for Smart City aspect Economy have chosen importance value of 1, will most probably chose the Smart City *definition* number 3 (Technological city).

The same procedure was performed for each combination of homogenized data and *Aspects*. From the achieved results, we can confirm that different view on a particular Smart City aspect has significant impact on the *definition* given respondents assigned for the Smart City concept.

6. Discussion

In this study, we focused on an essential topic – classification of discrete data. Due to its nature, such data usually require a large sample size to successfully determine dependencies among individual variables [25]. Due to the rather small sample size according to the available studies, logistic regression and clustering methods were not a viable tool in our case [4]. As an alternative, so-called lambda coefficient based on the conditional probability theory was used. Similarly to the results provided by [16], this study confirmed that lambda coefficient is indeed a viable method for the classification of data that could not be classified with other tools. This might be especially effective in a similar case to this study, where although ANN did provide very good classification results, we were unable to identify the underlying dependencies in data with the most commonly used tools like logistic regression, decision trees and cluster analysis. In that case, lambda might be the second step analysis that helps us to understand the data basis for the accuracy of the data classification by ANN. Performed investigation confirmed that in light of this methodology there are underlying dependencies in the data set that can be identified and analyzed. Thanks to the applied method, authors found that the choice of the Smart City *definition* among respondents is not entirely random, and there are certain characteristics and beliefs of the respondents that influence their understanding and interpretation of Smart City.

7. Conclusions

Although commonly used analysis tools failed to provide significant results, the lambda coefficient successfully delivered insight into dependencies in the analyzed data. Although not much can be said of statistical significance, the achieved results without doubt delivered an insight into inter-data relationships. The most prevailing definitions of Smart City among respondents are *Efficient city*, *City for people*, *Technological city* and *Sustainable city.* By means of conditional probability, we established that the importance evaluation of the *Aspect 4* variable (*Economy*) bears an information potential for the prediction of respondents definition of smart city. Although only four mentioned *definitions of SC* are possible to predict (due to the sufficient representation in respondent answers), their frequency presented in Tab. V proves that different rating of *Aspect 4* variable and has impact on the definition value. Future studies shall focus on broadening the application of lambda coefficient and possibly identifying it as a valuable step in variable identification and determination for more complex and profound analysis methods. As demonstrated in the presented study, lambda coefficient (λ) can be a valuable tool to understand and analyze prediction capabilities of various variables for models achieved with ANN. While neural networks can provide a strong classification tool, lambda coefficient might be a subsequent tool viable for understanding the basis of ANN accuracy rate.

Acknowledgement

The presented study is a result of the research done in the project “Smart City – Smart Region – Smart Community” – project (No. CZ.02.1.01/0.0/0.0/17_048/0007435) financed by Czech Operational Programme “Research, Development and Education” for the implementation of the European Social Fund (ESF) and the European Regional Development Fund (ERDF).

References

[1] DUDA M.D., NOBILE J.L. The fallacy of online surveys: No data are better than bad data. Human Dimensions of Wildlife, 2010, 15(1), pp. 55–64.

[2] ALWOSHEEL A., VAN CRANENBURGH S., CHORUS C.G. Is your dataset big enough? Sample size requirements when using artificial neural networks for discrete choice analysis. Journal of choice modelling, 2018, 28, pp. 167–182.

[3] FÖLDES D., CSISZÁR C., ZARKESHEV A. User expectations towards mobility services based on autonomous vehicle, In: 8th International Scientific Conference CMDTUR, 2018, pp. 7–14.

[4] DOLNICAR S., GRÜN B., LEISCH F., SCHMIDT K. Required sample sizes for data-driven market segmentation analyses in tourism. Journal of Travel Research, 2014, 53(3), pp. 296–306.

[5] FORMANN A.K. Die Latent-Class-Analyse: Einführung in die Theorie und Anwendung [Latent class analysis: Introduction to theory and application]. Weinheim: Beltz, 1984.

[6] SANTNER T.J., DUFFY D.E. The statistical analysis of discrete data. Springer Science Business Media, 2012.

[7] CHIPMAN H. Bayesian variable selection with related predictors. Canadian Journal of Statistics, 1996, 24(1), pp. 17–36.

[8] BENGIO Y., BENGIO S. Modeling high-dimensional discrete data with multi-layer neural networks. Advances in Neural Information Processing Systems, 2000, 12, pp. 400–406.

[9] NAGY I., SUZDALEVA E. On-line mixture-based alternative to logistic regression. Neural Network World, 2016, 26(5), p. 417.

[10] SAFAVIAN S.R., LANDGREBE D. A survey of decision tree classifier methodology. IEEE transactions on systems, man, and cybernetics, 1991, 21(3), pp. 660–674.

[11] MENARD S. Applied logistic regression analysis, Second Edition (Vol. 106). Sage, 2002.

[12] HARDY M.A. Regression with dummy variables (Vol. 93). Sage, 1993.

[13] XIE C., LU J., PARKANY E. Work Travel Mode Choice Modeling with Data Mining: Decision Trees and Neural Networks. Transportation Research Record, 2003, 1854, pp. 50–61.

[14] GOODMAN L.A., KRUSKAL W.H. Measures of association for cross classifications. Journal of the American Statistical Association, 1954, 49.268, pp. 732–764.

[15] VANIŠ M., URBANIEC K. Employing Bayesian networks and conditional probability functions for determining dependences in road traffic accidents data. In 2017 Smart City Symposium Prague (SCSP) (pp. 1–5). IEEE, 2017.

[16] TICHÝ L., CHYTRÝ M., SMARDA P. Evaluating the stability of the classification of community data. Ecography, 2011, 34(5), pp. 807–813.

[17] GOODMAN L.A., KRUSKAL W.H. Measures of association for cross classifications. Springer-Verlag, New York., 1979, pp. 2–34.

[18] DU K.L., SWAMY M.N.S. Recurrent neural networks. In Neural networks and statistical learning (pp. 351–371). Springer, London, 2019.

[19] LAU E.T., SUN L., YANG Q. Modelling, prediction and classification of student academic performance using artificial neural networks. SN Applied Sciences, 2019, 1(9), pp. 1–10.

[20] GOODFELLOW I., BENGIO Y., COURVILLE A. 6.5 Back-Propagation and Other Differentiation Algorithms. Deep Learning. MIT Press. 2016, pp. 200-–220. ISBN 9780262035613.

[21] PALIWAL M., KUMAR U.A. Neural networks and statistical techniques: A review of applications. Expert systems with applications, 2009, 36(1), pp. 2–17.

[22] NEDIC V., DESPOTOVIC D., CVETANOVIC S., DESPOTOVIC M., BABIC S. Comparison of classical statistical methods and artificial neural network in traffic noise prediction. Environmental Impact Assessment Review, 2014, 49, pp. 24–30.

[23] GOETHALS P.L., DEDECKER A.P., GABRIELS W., LEK S., DE PAUW N. Applications of artificial neural networks predicting macroinvertebrates in freshwaters. Aquatic Ecology, 2007, 41(3), pp. 491–508.

[24] KVÅLSETH T.O. Measuring association between nominal categorical variables: an alternative to the Goodman–Kruskal lambda. Journal of Applied Statistics, 2018, 45(6), pp. 1118–1132.

[25] BALDWIN S.A., FELLINGHAM G.W. Bayesian methods for the analysis of small sample multilevel data with a complex variance structure. Psychological Methods, 18, 2013, pp. 151–164.

Appendix – questionnaire

The questionnaire survey consists of 15 questions that can be divided into three parts, namely quality of life assessment (Q9-Q11), individual perception of Smart City (Q3-Q8) and socio-demographic questions (Q1, Q2 and Q12-Q15).

Q1 Where is your home located?

Q2 In what city do you live?

Q3 Which definition do you think best describes Smart City?

1. Ecological city – the city behaves with respect to the environment, tries to apply smart solutions that eliminate pollution and negative impacts on nature;

2. Efficient city – the city optimize costs and resources, applies smart solutions that improve efficiency, reduce costs and shorten duration;

3. Technological city – the city is developing according to the latest technological trends. There are technologies at every step in the city;

4. City for people – a city develops with respect to its citizens, according to their direct and indirect needs. The city applies smart solutions according to a long-term concept that was created together with them;

5. Sustainable city – the city applies solutions that ensure a quality environment for future generations;

6. Enabling the city – the city develops to give citizens comfort in the form of the freedom to do what they want;

7. Resilient city – the city is developing to increase resistance to negative influences and threats. It is installing smart solutions against floods, fires, crime, etc.;

8. Competitive city – a city that creates an environment for new ideas and innovations that will make the city economically prosperous and attractive for companies and people. The city focuses on smart solutions such as innovative HUBs, connects stakeholders, students and universities, etc.;

9. Transparent city – the city develops in data collecting and analysing, which serves better management of the city and citizens. The data are transparent.

Q4 Based on your opinion, what aspects are the most and the least important for living in the city? Please sort answers from 1 to 6, where 1 is the least and 6 the most important aspect.

1. Transportation;
2. People (Society);
3. Environment and energy;
4. Economy;
5. Buildings and public space;
6. Government.

Q5 Do you know examples of Smart cities in your country/worldwide?

Q6 Which 3 solutions from the menu you think Smart city should use?

1. Installation of various latest modern technologies;
2. Use of data. New information will be created thanks to the data;
3. Citizen participation. The approach of the office will change, people will be perceived by the city and the office as a customer;
4. Modern management and administration. The current costs of the city will be optimized;
5. Communication and information. The transparency of the city's activities and management will increase;
6. Intelligent urban transport. Transport from A to B will be faster, safer, multi-optional and more environmentally friendly;
7. Sustainable public spaces. The city will be safer;
8. Quality services (education, doctors ,...). The quality of life of the population will increase;
9. Smart buildings. Ecological buildings will be built in the city with regard to the future, the environment and the needs of the people;
10. Intelligent energy management system;
11. Use of renewable resources. The city will be more ecological;
12. Other (please specify).

Q7 How smart is city you live in? (where 1 represents lowest, 4 highest level of smartness)

1. Not at all;
2. Rather not;
3. Rather yes;
4. Absolutely.

Q8 What are the reasons for your previous answer?

1. I can use smart solutions that really help me and other citizens in everyday life;
2. I feel that the implemented smart solutions have helped to improve the quality of living;
3. I experienced some smart solutions during pilot projects I liked;
4. I know about many great smart solutions in my everyday life;
5. I don't like smart city plans for my city;
6. I don't understand the benefits of a smart city;
7. I don't know of any plans or solutions for our smart city;
8. I do not trust city government will fulfill promised visions of smart city;
9. I don't think smart solutions will help us because they are poorly chosen;
10. I feel uncomfortable living between technology.

Q9 Are you satisfied with quality of living in your city?

1. Dissatisfied;
2. Rather dissatisfied;
3. Rather satisfied;
4. Satisfied.

Q10 Please rate your satisfaction with the following aspects of quality of living in your city

Quality of living:

1. Safety;
2. Services at the office;
3. Job offers and job opportunities;
4. Shops, restaurants and cafes available;
5. Transport options (metro, trams, buses, bike, car, walk);

6. Activities and places to spend free time;
7. City management that I trust and am satisfied with its results;
8. Society me and my family live in;
9. Affordable and quality housing offer;
10. Public space around us (greenery, squares, streets, markets,...).

Rate:

1. Absolutely satisfied;
2. Rather satisfied;
3. Neither satisfied / dissatisfied;
4. Rather dissatisfied;
5. Completely dissatisfied.

Q11 Please now rate the same aspects of quality of living according to importance

Aspects of quality of living are the same as in question 10.

Rate:

1. Very important;
2. Rather important;
3. Neither important/ unimportant;
4. Rather unimportant;
5. Completely unimportant.

Q12 What gender do you identify as?

1. Male;
2. Female.

Q13 What is your age?

1. Less than 23;
2. 24-39;
3. 40-55;
4. 56-74;
5. Over 75.

Q14 What is your current employment status? You can choose more options

1. Employed Full-Time;
2. Employed Part-Time;
3. Unemployed;
4. Retired;
5. Private entrepreneur/ Self–employed;
6. Student;
7. On maternity / parental leave.

Q15 What is the highest degree or level of education you have completed?

1. High School;
2. Bachelor's Degree;
3. Master's Degree;
4. Ph.D. or higher;
5. Trade School.