UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS
UNECE Expert Meeting on Statistical Data Collection – Towards a New Normal?
26 to 28 October 2022, Rome, Italy

Session 2
19 September 2022

A machine learning based help desk approach for units involved in official surveys

Gianpiero Bianchi; Giampaola Bellini; Paola Bosso; Pasquale Papa (Istat)
gianbia@istat.it; bellini@istat.it; paola.bosso@istat.it; papa@istat.it

Abstract

The Italian National Statistical Institute (Istat) provides assistance services to respondents' requests for units involved in any kind of survey (enterprises, institutions, individuals), both directly and indirectly. In the latter case, service is provided by a specialized Contact Center. The assistance is guaranteed by synchronous (toll free number) and asynchronous channels (e-mail).

The centralized management of the services ensures homogeneity of treatment in requests for assistance and efficiency in controlling the process.

The assistance requests, also known as tickets, need to be classified by topic to be answered efficiently. The effectiveness of the ticketing system would greatly increase if this semantic categorization could be carried out automatically.

In this work, we pursue the objective of building an automatic process of assistance to the survey units by using natural language processing and machine learning techniques.

A great way to add value not only to respondents but also to those providing support is to work with a single point of contact. The basic idea is to route all communications through a single entity like a service desk.

Experiments on real-world data from the Contact Center of Istat confirm that an automatic ticket categorization is practically feasible and can reach a very good accuracy in very reasonable times.

A machine learning based help-desk approach for units involved in official surveys

G.Bellini, G.Bianchi, P. Bosso, P.Papa

ISTAT - Italian National Statistical Institute,

Directorate for Data Collection

keywords: Digitization, Automation, Classification, Machine Learning, Text Mining, Customer Support, Statistical Surveys

1. Introduction

The application of a centralized data collection model, such as the one adopted by Istat (Italian National Statistical Institute), involves the adoption of generalized solutions, of a methodological, technological and organizational nature, aimed at increasing the efficiency of the statistical processes, at reducing the resources employed, and controlling the burden on respondents [5].

The adoption of an optimal system and possible process automation solutions has important implications on the efficiency of the processes and can have important consequences on the Total Survey Error of the statistics produced, with particular regard to its non-sampling component [7].

Indeed, the analysis of the different types of respondents' problems can help to identify the critical areas for each survey. It makes it possible to identify the improvements to be made to the data collection processes.

Istat provides the assistance services to respondents' requests for units involved in any kind of survey (enterprises, institutions, individuals). The assistance is guaranteed by synchronous (toll free number) and asynchronous channels (dedicated email address).

A great way to add value not only to respondents but also to those providing support is to work with a single point of contact. The basic idea is to route all communications through a single entity like a service desk. The single point of contact model can be single or multi-channel. Users can use all kinds of channels like e-mail, phone, certified email (PEC), or others to communicate and request assistance. A single point of contact usually acts as a frontline in order to optimize the flow of tickets to other teams.

The assistance requests, also known as tickets, can be easily converted in short texts in natural language. They generally need to be grouped by topic to be answered efficiently. The effectiveness of the ticketing system would greatly increase if this semantic categorization could be done automatically: besides the obvious throughput boost and economic advantages, this will also provide a more consistent and unbiased subdivision. Ticket categorization problems arise in several different areas, and can be tackled by using machine learning techniques [2, 4].

In particular, we pursue this goal by building an automatic process of assistance to the survey units included in business surveys using text mining to extract the features of each ticket, and supervised classification algorithms to perform the categorization [1].

The adoption of an optimal approach as a single point of contact for the support requests that uses an accurate machine learning model for the classification of tickets could allow multiple benefits on different aspects of the Total Survey Error. To list some examples:

- simplifies the process of inquiry processing in order to speed up information flow and minimize ticket resolution times;
- optimizes human resources;

- improves the usability and functionality of data acquisition systems;
- offers a convenient way to coordinate and align information and requests in order to ensure consistency in communication;
- provides unified outputs to various types of requests and removes uncertainty of whom to contact with various types of requests;
- eliminates the risk of miscommunication and wrong interpretation of tasks.

The proposed methodology works at the formal level using a data-driven approach, and thus it could also be applied to the semantic categorization of other texts with different origins or in different languages.

2. Inbound Service and digital service management

In order to automate the management of customer services and collect uniform data, the first step is to digitize the service delivery processes on a digital platform. It is necessary to map the processes and needs of the people involved, identify recurring activities with low added value and review workflows in order to simplify the request for services, both on the user side and on the back-office side.

The standardization of digital processes facilitates the collection of already structured data and allows for a single point of control in which to aggregate information on the people served by multiple channels, increasing the omnichannel capabilities.

The second step is to automate repetitive tasks and increase self-service by means of:

- the configuration and training of automatic classification models with the analysis of text and unstructured data;
- the automation of easy to solve tickets, in the presence of recurring problems;
- the automated opening of tickets in the most complex cases, accompanied by data to speed up the intervention.

The third step is to improve the quality and perceived innovation of the service by means of:

- the collection of feedback;
- satisfaction indicators;
- proactive problem solving prior to the request.

In the specific case of Istat, the Inbound service provides assistance and support to responding units in the access and navigation of the data acquisition systems (i.e. Business statistical Portal for enterprises), as well as on the general rules that define the statistical activity and on the legal obligations for respondents. Finally, it provides answers to the most recurring questions about major instances of the survey's content [6].

The assistance is guaranteed by synchronous (toll free number) and asynchronous channels (dedicated email address).

The infrastructure on which all the system is based is the tool called shared agenda, that gives the possibility to share the ticket generated by each incoming inquiry at the correct destination.

In fact, there are two levels of assistance: first-level assistance is the one provided directly by the CC operators to solve the most recurring problems managed using FAQs, while the second one refers to cases with a higher degree of complexity that recurrently involves both DCI non-thematic and thematic experts. For requests that are not solvable directly by the CC, a tool - the shared agenda - presenting features useful for managing and sharing the received instances is used.

It becomes evident that a set of FAQs has to be provided to the CC operators aimed at ensuring the uniformity of the unit treatment by using a set of harmonised answers in both services. The set of FAQs are of two kind: one solves inquiries on use of Business statistical Portal, another one is survey specific and tackles different

kinds of issues. This kind of organization implies that – particularly for structure survey on enterprises – a specific training has to be supplied every time a specific survey has launched.

The volume of assistance necessary to answer units involved in the surveys during the year can be of course planned in advance, but since there are peaks of activities notably when surveys are launched or closed and when census are run, sometimes it is difficult to organize the assistance precisely according to the workload realized in that specific period of time. In this period the overall services shows a decrease in the general efficiency, as the operators offering assistance are overloaded. A set of indicators named SLA (Service Level Agreement) allows measuring and monitoring the quality and the efficiency of the service over time.

Thus, a way to increase efficiency in general and especially in these critical moments is to invest in automation of specific part of the process.

3. **Methods**

In the literature, the problem of introducing automatic solutions for the semantic categorization of tickets has been approached with a variety of techniques, more and more switching to deep learning in recent times. Quite often, the overall strategy requires a first phase to extract from the text of each ticket the features describing that ticket, and then another phase to determine the partitioning. The first phase generally use text mining, which is the branch of data mining concerning the process of deriving high-quality information from texts. In our case, the aim is describing each free-form textual ticket with a standardized data record [3]. On the other hand, the second phase can be viewed as a classification operation. Classification is the supervised process that takes a training set of elements, each of which labeled with a class value, and learns a criterion to predict the class label of other unseen elements.

In our case, it assigns a subject (taken among a set of predefined subjects) to each data record representing a ticket. Note that, if the set of subjects is not predefined, this part can be seen as clustering, instead.

However, it is a difficult task, since it requires an automatic extraction of the meaning of the text, followed by a classification, which is often multiclass. These text messages often are the result of a summary made by an operator. In many cases, they are very short and sometimes incomplete texts.

In many cases, for example requests received by telephone and recorded, there is no metadata supporting the semantic categorization. In other cases, such as emails, the metadata (e.g., the subject of the email) may be present, but they are often too generic, not reliable enough, or not concerning the desired categorization. In all the above cases tickets need to be categorized homogeneously and preferably automatically. Evidently, basic features of the tickets, like length of the message or similar measures, are not meaningful for the semantic categorization. Even basic presence or absence of predetermined words is not enough, because in many cases similar words (answer, form, compile, fill, etc.) may describe completely different types of problems. Therefore, a more in-depth analysis is needed to extract the significance of a ticket from its text in natural language.

Therefore, it is necessary to introduce automatic solutions capable of performing frequent and high-volume manual tasks and extracting hidden relationships and patterns.

A scheme of the overall procedure described in this work is reported in the subsequent Figure 1. After an initial Tokenization (individuation of the words within the sequence of characters) and Stop-words Elimination (elimination of useless parts, like articles, etc.), we perform Lemmatization with Part-Of-Speech recognition. This means that, for each word, we remove the inflectional ending to identify its basic lemma. This allows to recognize together the different inflected forms of a word (e.g., plurals of nouns, tenses of the verbs, etc.).

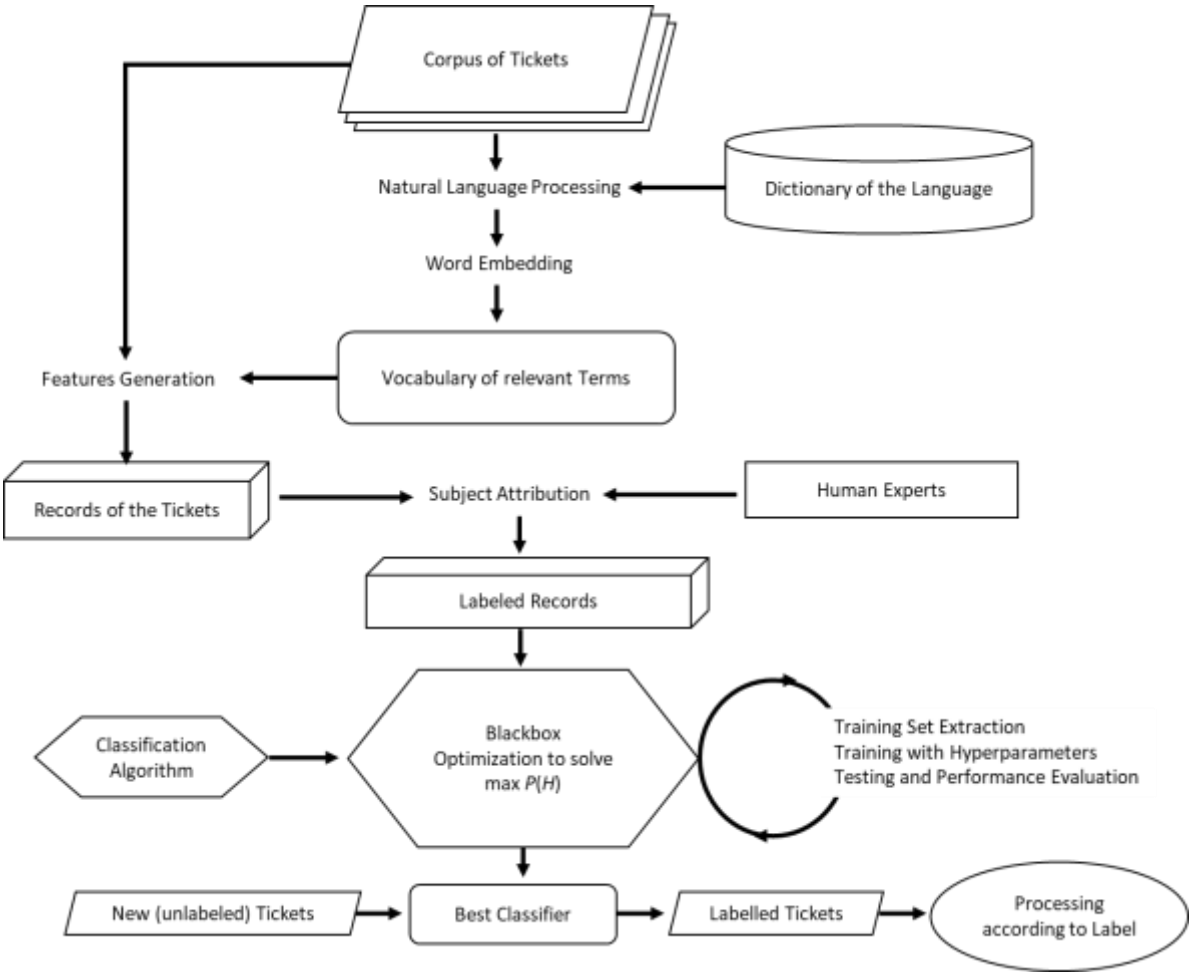

Figure 1: Overall scheme of the proposed approach

Moreover, we still keep track of which part of speech each word is (e.g., noun, verb, adjective, etc.). This Natural Language Processing (NLP) is done by using Gensim python library from scikit learn. Since the tickets are in Italian language, we perform the above operations by using an Italian dictionary. However, the language can easily be changed by simply switching the underlying dictionary.

After this, we need to convert each ticket into a data record constituting a "standardized description" of the ticket. This feature extraction is done by using the word embedding algorithm Word2vec. This algorithm uses a shallow neural network to learn word associations in a large corpus of text, all the tickets in our case. Hence, it can detect synonymous words. In particular, Word2vec represents each distinct word with a particular list of numbers called a vector. The vectors are generated in such a way that the cosine similarity between the vectors indicates the level of semantic similarity between the corresponding words.

In more detail, we have assembled a corpus composed of the of text of 15,000 real tickets received by the Contact Center of Istat. After the above described NLP steps, we identified in the word embedding operations a vocabulary with 31,000 most relevant lemmas, later decreased to 424 by cutting words not enough frequent. By projecting on this set of lemmas, each ticket has been converted into a vector with 424 elements, each of which computed as the frequency of each relevant lemma in that ticket (number of occurrences normalized by the size of the ticket).

After obtaining the training set of labeled tickets, we set up a classification phase, which is able to predict the class of unlabeled tickets by learning the classification criteria from the above training set. We perform experiments with both deep and traditional learning strategies.

Our approach is formal and data driven: all the relevant information is extracted from the data, except for the class labels of the training set which are of course externally defined. Hence, the proposed approach is not confined to the specific context described above, but can be adapted to the categorization of other texts with different origins.

Moreover, we propose an innovative technique for the determination of the hyperparameters of the classifiers. The classification algorithm needs a suitable hyperparameter configuration to produce a practically useful categorization. As recently highlighted by several researchers, the selection of these hyperparameters is often the weak link in many machine learning applications.

In general, to use each classifier, a tuple of values $V = (v_1, ..., v_n)$ must be assigned to a set H of hyperparameters. Note that, for simplicity, we call here "hyperparameters" all the values needed by the classifier which are not learned by data, not distinguishing between the categories of those values. On the contrary, the values learned by data (e.g., neuron weights in a neural network) are generally called only "parameters" and will not be set in our hyperparameters optimization but will be determined during the training of the classifier, which constitutes an "inner loop" of our procedure (see also Figure 1).

Generally speaking, the majority of the hyperparameters are bounded to take integer values (e.g., the number of neurons in a neural network), even if some can vary with continuity within a given interval (e.g., the kernel coefficient γ in a support vector machine), and some are categorical (e.g., the choice of the activation function of a neuron). Thus, each value v_i must belong to its feasible domain D_i, and consequently we can define the set H of all the feasible tuples of hyperparameter values. The choice of a tuple of values, that is, the choice of a point in $\mathcal{H}$, determines the behavior of the classifier, and so it may dramatically affect its prediction performance P. As a matter of fact, small variations in hyperparameter values may sometimes determine a huge variation in P. Therefore, the values assigned to the hyperparameters must be selected very carefully, and they cannot simply be "default" values or similar.

We propose to view the hyperparameter choice as a higher-level optimization problem, where the hyperparameters H are the decision variables and the objective is the performance P of the classifier, evaluated by choosing a performance measure appropriate to the specific classification case.

$$\max_{H \in \mathcal{H}} P(H) \tag{1}$$

In our case, the accuracy (defined as the overall percentage of correct class predictions over the test set) was deemed to be the appropriate measure, even if in other cases one may prefer precision, sensitivity, F-1 score, etc. However, the above problem has a main difficulty. Even though P clearly depends on H, this dependency cannot be expressed in analytical form.

To solve it despite this issue, we propose to adopt a black-box optimization approach and to use a derivative-free algorithm. Similar algorithms do not need the explicit optimization model, in particular the analytic expression of the objective function; they only need to numerically evaluate the objective function over a number of points (i.e., tuples of values of the hyperparameters). Due to the discrete nature of the choices, we use for this task a recently proposed algorithm [8] based on primitive directions and nonmonotone line search which is able to deal with integer decision variables.

Due to the time required by each hyperparameter evaluation (full training and testing with 5-fold cross validation), we could experiment only the sets of values for the hyperparameter that were deemed sufficiently promising. However, in principle, our technique can handle any sets of hyperparameter values, provided that the required computations can be carried out in practice. Furthermore, since the number of

points tested in a grid search tends to be exponential (all the possible combinations), for mere computational reasons we have been occasionally forced to reduce in the grid search the number of values tested for each hyperparameters with respect to those considered in the optimization approach.

In more detail, we build a CNN with three layers of type 2D convolutional, each followed by a 2D Max Pooling layer, and one Dense layer in output. We choose this architecture as a good compromise between speed and performance, since the solution of problems 1 requires to train and execute the network a large number of times with many different sets of hyperparameter values.

The hyperparameter tested for this CNN are:

- Embedding_dim: width of the kernel matrix for the 2D convolution window.
- Filter_sizes: the number of words we want our convolutional filters to cover.
- Num_filters: the number of output filters in the convolution.
- Optimizer: the optimization technique used in the gradient descent when training the network.
- Loss_function: the function used to calculate the error when training the network.
- Activation_conv: activation function for the convolutional layers.
- Activation_dense: activation function for the final dense layer.
- Epochs: an epoch is one complete pass through the training data.
- Class_weights: specifies how the errors in the different classes are weighed in the loss function.

Hence, the main methodological contributions of this work include:

1) A methodology to solve the difficult multiclass classification problem of the categorization of tickets written in natural language, based on text mining and classification, which works at the formal level using as much as possible a data-driven approach. Thus, it could also be applied to the semantic categorization of other texts with different origins or in different languages.

2) An approach to the difficult problem of determining the hyperparameter configuration of a generic machine learning procedure. Again, this approach is purely formal, hence it can solve other problems of hyperparameter optimization, also for hyperparameters assuming integer or even categorical values.

4. Experimental results

We design our experiments in order to determine the effectiveness of an automatic approach for ticket categorization problems, given the difficulties of a real world case, and to compare the black-box approach proposed for the optimization of the hyperparameters to a standard grid search, which is very often used in the literature. A grid search is the simple evaluation of all the combinations of the possible hyperparameter values. This corresponds to a complete enumeration approach in the solution of problem 1. If not interrupted, it clearly guarantees completeness, hence in theory it always reaches the optimal solution, however this may require in practice very long times, often excessive.

We considered both the case of the tickets received from a Contact Center, that is run by a private specialised company, and the requests received through ordinary e-mails and certified e-mails, managed directly by Istat.

The Contact Center currently manages about 60 statistical investigations and processes an average volume of about 80,000 requests per year.

Instead, as regards the channel of ordinary e-mails and certified e-mails, received in the field of economic surveys, there are up to over 30,000 requests for assistance per year.

We have focused on the tickets arising from economic surveys with the aim of defining a training set. In particular, the representative FAQ of the various cases of requests for assistance from respondents were considered as training set.

This type of survey contains many of the critical issues that users may encounter during the compilation, due to the complexity of the questionnaires. Indeed, they have the following characteristics: they require detailed quantitative information; they use tabular forms for filling out many questions; they contain a large number of prompts, many of which are blocking (also known as hard prompts).

The ticket categorization operation is currently performed by human experts, and although it is often performed in a more coarse-grained manner, it requires a considerable amount of work.

To perform the classification task, we have prepared a training set of 8,000 tickets, whose class has been determined by experts of the field using 7 distinct classes, which are listed below:

1. requests for general information on the survey;
2. problems relating to the usability of the electronic questionnaire;
3. requests for assistance for a specific question in the questionnaire;
4. criticalities in the communication process between Istat and end users;
5. obligation to participate in the investigation;
6. requests with ambiguous content;
7. requests with multiple problems highlighted at the same time.

The class subdivision of the mentioned training set is the following: 3% in class 1, 3% in class 2, 9% in class 3, 34% in class 4, 15% in class 5, 5% in class 6, 31% in class 7.

To perform the classification task, we have selected 50% of the total dataset as training set, and the rest has been used for test. The extraction has been randomly performed 5 times, and all performance results are averaged on the 5 trials. The resulting 7-class problem is not trivial. Just as an example, preliminary tests with some “default” hyperparameter values obtain less than 50% in accuracy.

As reported in Table 1, by considering all the values of the hyperparameters described in the previous Section, in the case of CNN we obtain 7,680 possible hyperparameter configurations for the black-box optimization approach, and 240 hyperparameter configurations for the grid search approach.

However, only a small fraction of the points of this search space (212) is actually evaluated by the optimization approach (Evaluated points in Table 1), which is therefore able to find a better solution in less time than the grid search.

	CNN	
	Black box	Grid search
Hyperparameter configurations	7,680	240
Evaluated points	212	240
Enumeration percentage	2.76%	100%
Time in sec.	318,500	375,000
Solution accuracy	89.92%	86.15%

Table 1. Performance evaluation of the proposed optimization approach

In the case of our 7-class problem, the experiments on the classification of tickets - whose real class was known - show that the proposed approach allows to build a classification procedure able to obtain about 90% in accuracy.

Finally, to further test our approach, we have applied this procedure on a large dataset of 193,419 unlabeled tickets arising from several surveys of economic scope in the period 2016-2019.

As hyperparameter values we have considered the solution of optimization problem 1.

The overall classification time of all tickets was about 170,000 seconds. Although the real labels of these tickets are not available, and so the exact accuracy is not computable over this dataset, the results have been judged very satisfactory from a practical point of view, and show that an automatic treatment of these tickets is feasible and useful.

5. Conclusions

The building of an automatic process of assistance to the survey units first requires the introduction of an appropriate digital service management strategy able to guarantee:

- Standardization of digital processes.
- Automation of repetitive tasks.
- Improvement of the quality and perceived innovation of the service.

Subsequently, we need to introduce automatic solutions for the semantic categorization of tickets in order to increase the efficiency and effectiveness of the ticketing system.

In particular, the semantic categorization of tickets received by a contact or costumer center is an important practical problem. Its solution can provide several advantages, improving both the efficiency of the ticketing system and the quality of the answers, and it may also help in designing or improving the procedures relating to the tickets. However, it is a difficult task, since it requires an automatic extraction of the meaning of the text, followed by a classification, which is often multiclass. This work proposes an approach to this problem, based on text mining and classification, which can use either deep or traditional learning algorithms.

The proposed methodology works at the formal level using a data-driven approach, and thus it could also be applied to the semantic categorization of other texts with different origins or in different languages. Experiments on real-world data from the Contact Center of Istat and from dedicated email addresses confirm that an automatic ticket categorization is practically feasible and very useful. In particular, experiments on the classification of tickets whose real class was known, show that the proposed approach can reach a very good accuracy in very reasonable times.

Future work should first include the experimentation of some technological solutions able to implement a digital platform that works as a single point of contact. On this platform information on the people served by multiple channels is aggregated, increasing the omnichannel capabilities. The basic idea is to route all communications through a single entity like a service desk.

In addition, we must extend the analysis made for the semantic classification of the Contact Center tickets to all communication channels with respondents.

References

[1] Bellini G., Bianchi G., Di Paolo GG., Papa P. *Towards a selective automation process of assistance to the survey units included in business surveys*, 2022 Business Data Collection Methods Workshop, Session T1 Language processing and machine learning techniques, Statistics Norway, Oslo, 13 - 15 June 2022.

[2] Bianchi G., Bruni R. *Website categorization: A formal approach and robustness analysis in the case of e-commerce detection*. EXPERT SYSTEMS WITH APPLICATIONS, vol. 142, ISSN: 0957-4174, doi: 10.1016/j.eswa.2019.113001.

[3] Bianchi G., Bruni R., Scalfati F. *Identifying e-Commerce in Enterprises by means of Text Mining and Classification Algorithms*. MATHEMATICAL PROBLEMS IN ENGINEERING, vol. 2018, p. 1-8, ISSN: 1024-123X, doi: 10.1155/2018/7231920.

[4] Bianchi G., Bruni R. *Effective Classification using a small Training Set based on Discretization and Statistical Analysis*. IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, vol. 27, p. 2349-2361, ISSN: 1041-4347, doi: 10.1109/TKDE.2015.2416727.

[5] Bellini G., Monetti F., Papa P. *The impact of a centralized data collection approach on response rates of economic surveys and data quality: the Istat experience*. STATISTICS AND ECONOMY JOURNAL VOL. 100 (1) 2020 ISSN 1804-8765 (Online) *ISSN 0322-788X (Print).*

[6] Bellini G., P. Bosso, S. Binci, S. Curatolo, F. Monetti *Centralized Inbound and Outbound Contact center Service as New Strategy in Data Collection.* Fifth International Workshop on Business Data Collection Methodology. Lisbon, 19-21 September 2018. Available at https://www.ine.pt/scripts/bdcm/doc/00_BDCMLisbon2018_BP.pdf pgg 138-141.

[7] Groves R.M. and Lyberg L. (2010), Total Survey Error: Past, Present, and Future Public OpinionQuarterly, Volume 74, Issue 5, 2010, Pages 849–879, https://doi.org/10.1093/poq/nfq065.

[8] Liuzzi, G., Lucidi, S., Rinaldi, F. An algorithmic framework based on primitive directions and nonmonotone line searches for black-box optimization problems with integer variables. Mathematical Programming Computation 12, 673-702.

[9] Salemink I., Dufour S., Van Der Steen M. (2019). Vision Paper on Future Advanced Data Collection, *UNECE Statistical Data Collection Workshop 'New Sources and New Technologies'*, Geneva 14-16 October 2019.