Wiadomości Statystyczne. The Polish Statistician, 2023, *68*(10), 49–70 https://doi.org/10.59139/ws.2023.10.3
Spisy powszechne – problemy i wyzwania / Issues and challenges in census taking
Zgłoszony/submitted: 5.04.2023, poprawiony/revised: 3.08.2023, zaakceptowany/accepted: 16.08.2023

Digital population and housing census – the experience of Serbia

Miladin Kovačević,[a] Mira Nikić,[b] Branko Josipović,[c] Snežana Lakčević,[d] Vesna Pantelić,[e] Nevena Mitrović,[f] Adil Kolaković,[g] Petar Korović[h]

Abstract. The aim of the paper is to present the experience of the Republic of Serbia in conducting the 2022 Census of Population, Households and Dwellings, focusing on the employment, legal framework and financing of the census as well as on its successful implementation. It discusses strategic decisions on data collection and the integration of information technology – including geospatial data, data collection techniques, machine learning, record linkage and monitoring system – to overcome the challenges posed by the census. The paper addresses the census undercoverage, explores the use of administrative data for item imputation, and examines the development of a statistical population register. The study demonstrates the benefits of adopting a digital-census approach: significant improvement of accuracy, cost reduction and acquired expeditiousness.

The Statistical Office of the Republic of Serbia conducted a digital census combined with traditional methods, excluding self-enumeration, along with the use of administrative data for item imputation, and recommends this approach as the most effective way to obtain precise and comprehensive information about a population, including its demographic characteristics, geographic distribution and overall size.

Keywords: 2022 Census of Population, Households and Dwellings, digital census, geospatial data, monitoring system, machine learning, administrative data, record linkage, imputation, statistical population register, Serbia

JEL: J18, M15, N34

[a] Statistical Office of the Republic of Serbia, Serbia. ORCID: https://orcid.org/0009-0002-2398-7955. E-mail: miladin.kovacevic@stat.gov.rs.
[b] Statistical Office of the Republic of Serbia, Serbia. ORCID: https://orcid.org/0009-0003-2340-8789. Autor korespondencyjny / Corresponding author, e-mail: mira.nikic@stat.gov.rs.
[c] Statistical Office of the Republic of Serbia, Serbia. ORCID: https://orcid.org/0009-0009-3191-5784. E-mail: branko.josipovic@stat.gov.rs.
[d] Statistical Office of the Republic of Serbia, Serbia. ORCID: https://orcid.org/0009-0002-8106-8412. E-mail: snezana.lakcevic@stat.gov.rs.
[e] Statistical Office of the Republic of Serbia, Serbia. ORCID: https://orcid.org/0009-0002-6430-9288. E-mail: vesna.pantelic@stat.gov.rs.
[f] Statistical Office of the Republic of Serbia, Serbia. ORCID: https://orcid.org/0009-0000-2730-8011. E-mail: nevena.mitrovic@stat.gov.rs.
[g] Statistical Office of the Republic of Serbia, Serbia. ORCID: https://orcid.org/0009-0008-2553-3707. E-mail: adil.kolakovic@stat.gov.rs.
[h] Statistical Office of the Republic of Serbia, Serbia. ORCID: https://orcid.org/0009-0004-5761-8554. E-mail: petar.korovic@stat.gov.rs.

Cyfrowy powszechny spis ludności i mieszkań – przykład Serbii

Streszczenie. Celem artykułu jest przedstawienie doświadczeń Republiki Serbii w zakresie organizacji Powszechnego Spisu Ludności, Gospodarstw Domowych i Mieszkań 2022, ze szczególnym uwzględnieniem zagadnień dotyczących zatrudnienia personelu, ram prawnych i finansowania tego badania oraz warunków jego udanej realizacji. Praca skupia się na strategicznych decyzjach w sprawie zbierania danych oraz zastosowania technik informatycznych, takich jak: wykorzystanie danych przestrzennych, cyfrowe metody uzyskiwania danych, uczenie maszynowe, łączenie rekordów czy system monitorujący, mających na celu sprostanie wyzwaniom związanym ze spisem. Autorzy poruszają także kwestie niedostatecznego pokrycia spisu oraz wykorzystania rejestrów administracyjnych do imputacji danych. Ponadto poświęcają uwagę opracowaniu i udoskonalaniu statystycznej ewidencji ludności, dokładności danych, obniżeniu kosztów i zwiększeniu efektywności badania.

Główny Urząd Statystyczny Republiki Serbii przeprowadził spis powszechny w sposób cyfrowy, łącząc ten mechanizm z metodami tradycyjnymi (z wyłączeniem samospisu) i posiłkując się rejestrami administracyjnymi w celu imputacji danych. Metoda ta jest w artykule rekomendowana jako najefektywniejszy sposób uzyskania precyzyjnych i wyczerpujących informacji na temat populacji, w tym jej charakterystyki demograficznej, rozmieszczenia przestrzennego i liczebności.

Słowa kluczowe: Powszechny Spis Ludności, Gospodarstw Domowych i Mieszkań 2022, spis cyfrowy, dane geoprzestrzenne, system monitorujący, uczenie maszynowe, dane administracyjne, łączenie rekordów, imputacja, statystyczna ewidencja ludności, Serbia

1. Introduction

The Serbian census has a long history dating back to 1834 with nearly five-year intervals until the First World War. The 1866 Census was the first modern and complete survey of this kind. Between the world wars, only two censuses were conducted. After the Second World War, a census was urgently conducted in 1948 to collect data on damages caused by the war. Starting from 1961, a ten-year periodicity for censuses was established. Thus, the subsequent censuses were conducted in 1961, 1971, 1981 and 1991. The census of 2002 was delayed due to financial constraints and the authorities in Montenegro requesting postponement. The 2011 Census returned to the established periodicity, and the latest survey of this kind was conducted in October 2022, having been postponed due to the COVID-19 pandemic. According to the most recent results, the population of the Republic of Serbia totalled 6,647,003 usual residents, 2,589,344 households and 3,613,352 dwellings.

Serbia's censuses have been regulated by special legal provisions since 1884. Nowadays, as an EU candidate, the country follows the EU census legislation, which

upholds diversity and stipulates that countries should define the content of their census legislation in such a way that it is compliant with the national legal practices and procedures. The Law on the 2021 Census of Population, Households and Dwellings regulates the preparation, financing, organisation and implementation of the Census in the Republic of Serbia. It primarily determines the period and method of the census implementation, with the Statistical Office of the Republic of Serbia (SORS) playing the key role in the process. It also specifies the tasks of census commissions, formed for the purpose of conducting the survey, as well as the competences and obligations of ministries and other governmental bodies and organisations involved in the preparation and implementation of the survey. The law moreover stipulates the obligations of the participants, persons who directly perform duties related to the census, as well as the use and protection of the census data, the publication of the results and the financing of the survey.

The SORS went through all the procedural steps for the adoption of the census law (the public debate, the analysis of the probable effects of the regulation, collecting opinions of the relevant state authorities, etc.). Also, the census law was enhanced and aligned with the Law on Personal Data Protection, which ensured that all the aspects of the preparation and implementation of the census were acceptable to the public.

For the first time in history, the text of the law together with the Table of Concordance (where the concrete articles were compared to the relevant articles from the EU legislation) were submitted to the European Commission. In response, the European Commission suggested certain methodological improvements, but in general, the law was assessed as compliant with European legislation.

In April 2021, the above-mentioned law was amended by the Law on Amendments to the Law on the Census of Population, Households and Dwellings, which sanctioned the postponement of the census due to the COVID-19 pandemic.

The aim of the paper is to present the experience of the Republic of Serbia in conducting the 2022 Census of Population, Households and Dwellings, focusing on the employment, legal framework and financing of the census as well as on its successful implementation. It examines the undertaken strategic decisions regarding data collection and the integration of information technology (such as geospatial data, data-collection techniques, machine learning and the monitoring system), emphasising their role in addressing the challenges posed by the census. The paper also addresses the issue of undercoverage and explores the use of administrative data for item imputation. Future-oriented attitude to the topic in question adopted in the paper invites discussion on the development of a statistical population register.

The overarching objective of the paper is to highlight how the digital-census approach, in combination with the use of administrative data for item imputation (without self-enumeration), can lead to the improved accuracy, reduced costs and a faster process of data collection, processing, and dissemination.

2. Financing and implementation of the Population Census in Serbia

The population census was a part of the 'EU for Development Statistics in Serbia' project, funded by the EU and the Republic of Serbia through the Pre-Accession Assistance (IPA) 2018 national programme. The grant contract was signed on 29th July 2019. The cost of the census was estimated at approximately 21.2 million euro excluding IT equipment, and 29.6 million euro including it. The government of Serbia committed itself to providing co-financing (32.77%) and funds for the purchase of the IT equipment. The cost of the census, compared with 2011 census, increased due to changes in tax laws. They resulted in the additional cost of 4.6 million euro for taxes and contributions for enumerators and trainers, and 5.5 million euro for all externally-engaged field workers.

The budget for the Population Census was planned in 2017, with an average enumerator salary of 45,000 RSD (370 EUR), equivalent to the average monthly earnings in Serbia. The inflation rate grew significantly between the previous census and 2021, i.e. by 139.2% in the period from 2012 to 2021.

The SORS decided to postpone the Population Census due to the COVID-19 pandemic which posed a risk to the preliminary activities and fieldwork operations. Field activities were initially postponed for six months and then for an additional year, so finally the census was scheduled for October 2022 (instead of April 2021). This delay necessitated additional expenses that could not have been initially planned for.

Due to the COVID-19 pandemic, additional funds were required for the census-related activities, as well as for the delayed field work. The European Commission rejected the request for the increase of the funds, and consequently, the Serbian Ministry of Finance was asked to provide additional resources for remuneration of the enumerators and instructors. The project budget did not account for inflation, which resulted in lower remuneration for people who directly served the census. This could have led to difficulties in finding qualified staff, as 18,000 people to be recruited needed to be computer-literate to participate in the census.

The SORS received confirmation from the Ministry of Finance in March 2022 that they would provide the necessary funds, meeting an important prerequisite for the successful census. The additional funds increased the share of the Republic of Serbia to 44.83%, while the EU grant amount remained the same, which effectively

decreased the EU share to 55.17%. A detailed Financial Instructions document set out the principles for the use of funds, and the planned financial activities proceeded without any significant problems.

3. Strategic decisions regarding data collection in the 2022 Census

After the completion of the 2011 Census, conducted by means of paper questionnaire and OCR (optical character recognition) technology used to scan paper questionnaires and input the data into a database, the SORS began examining new approaches to a traditional census. The basic aim was to explore the possibilities for reducing costs, minimising burden for respondents, and improving the availability of more up-to-date estimates and surveys in the interim period.

Following international trends and examples of good practices, the possibility and suitability of using administrative sources for the census purposes were analysed in detail. It was concluded that there were still no conditions for the implementation of a fully register-based census. Considering that different registers are continuously developing, some data can be used in different phases of the preparation and implementation of the census (for defining boundaries and creating maps of enumeration areas, controlling the coverage of individual contingents, imputations, etc.).

Activities connected to the implementation of the 2019 Pilot Census were aimed at testing all methodological, technical and IT solutions. Apart from using two new data collection techniques, i.e. computer-assisted interviews (CAPI) and self-enumeration via Internet (CAWI), the monitoring system as well as the methods for the collection and linkage of census and geospatial data were examined. Although the phases of preparation and implementation are considered to have been successfully done, the response rate for self-enumeration via Internet was very low and the quality of the obtained data was questionable. Therefore, it was decided that self-enumeration via Internet should not be used as a method of collecting data in the 2022 Census.

After an additional analysis of all the pros and cons of certain methods of the census implementation, it was concluded that the transition to a census based on administrative sources would be a complex and long process. For this reason, the 2022 Census had to be a traditional one, yet it was modernised by using the CAPI method, i.e. laptops for face-to-face interviews. The application of the above-mentioned method is considered a significant modernisation of the census processes; it increases data accuracy and speeds up data processing. It also significantly reduces the number of the engaged enumerators and makes it possible

to monitor the census activities in real time, as well as starting the dissemination of the final results faster than while using the traditional method of data collection.

Following the announcement of the COVID-19 pandemic, the SORS faced new challenges related to conducting the census and developing actions or options for reducing the impact of the epidemic. Primarily, the census field implementation was postponed from April 2021 to October 2022. Afterwards, the SORS focused on developing various scenarios, such as: the potential adjustments of census questionnaires, the adoption of additional data collection method, and changing methods and plans for the enumeration of special population groups.

Given that some groups of the population, e.g. those living in special institutions[1], were identified as potentially exposed to a higher risk of the COVID-19 infection, their enumeration was carried out by the employees of this institutions through a web application (CAWI).

Apart from CAPI, in the cases where face-to-face interview was not possible due to the unfavourable epidemiological situation, enumeration was performed via telephone (CATI). Likewise, the census activities were redesigned to respond to the effects of the pandemic and its impact on the quality of the census output.

4. Geospatial technology in support of the census operations

Geospatial technology during the enumeration was a fundamental component of the Serbian Census 2022. The SORS implemented recommendations from the UN Guidelines concerning the use of electronic data collection technologies in population and housing censuses (United Nations, 2019). In recent years, the Geodetic Authority of the Republic of Serbia has developed and maintained a sophisticated national geospatial database that consists of locations of housing units, addresses, administrative boundaries, statistical boundaries, and a digital map of all the streets in the country. In preparation for the census, the Geodetic Authority provided the SORS with a digital orthophoto produced from the aerial photogrammetric data, and a national geospatial database with the aforementioned variables.

During the census, geospatial technology was used to support field operations by providing field maps and address lists. Prior to the census, it was used to design and digitise enumeration areas, establish an address-coding system, predetermine coverage areas and prepare maps. During the enumeration, geospatial technology was also used to evaluate data coverage and quality.

[1] Institutions for the execution of criminal sanctions (prisons, correctional institutions), social welfare institutions (homes for the elderly, children with disabilities, etc.), special hospitals for psychiatric diseases.

5. Data collection and transfer

5.1. Considerations for selecting hand-held devices

The selection of hand-held devices for data collection in the Census 2022 was based on three key parameters: operating system selection, device characteristics and price. The decision to use a specific operating system (OS) was made after considering various factors such as the IT infrastructure, business applications, security, human resources and the solution already developed and implemented at the SORS for the CAPI surveys. Under the chosen OS, three types of devices could be used: a laptop, a '2 in 1' device (a portable computer that had features of both a tablet and a laptop), and a tablet. The choice of device was based on several factors such as the availability, usability in 2021 and beyond, price within the budget, resistance to physical damage, screen size, weight, resistance to market fluctuations, storage, keyboard, portability, processing power and external ports. A laptop with the chosen OS was selected, as it fitted the procurement budget and met all the necessary requirements. The SORS acquired 15,500 laptops with similar configurations and weight less than 1.5 kg a piece. In addition, the SORS insisted on a service-contracted agreement with the vendors for defective laptop replacements within 24 hours, and leased warehouses for storing and preparing equipment for use and distribution.

5.2. Data collection using laptops (CAPI)

The SORS developed its own metadata-based data integration platform (IST), which offered various data collection methods, including CATI, CAPI and CAWI. The IST platform used a relational database format for data storage (LocalDB for CAPI and MS SQL Server or SQL as service for other data-collection modules) and encryption to ensure data security. The IST platform allowed enumerators to work offline on laptops and to synchronise data when online.

During the census, the IST platform organised the questionnaire into four segments: building, dwelling, household and personal information. Enumerators received through their laptops lists of addresses that they had to visit. The application directed the enumerator through each part of the questionnaire during data entry and prevented them from skipping any segments. However, in the data-editing phase, they were allowed to skip segments to immediately approach preferred fields rather than navigating through all the questionnaires. The IST platform was non-linear, which allowed enumerators to directly enter the dwelling or personal information sections of the questionnaire and make any necessary changes.

To ensure accurate data entry, the IST platform sent warning and stop messages when users entered invalid data or skipped the required questions. This improved data quality and reduced manual cleaning. Skip patterns were used in order to allow certain questions or sections of a form to be skipped based on previous answers. That made the survey more efficient. Data validation rules alerted enumerators to errors in the entered data, helping to ensure the sufficient quality of data.

To guarantee data security, the information stored on laptops and sent to the server was encrypted with a key or password. Encryption obfuscates data, rendering it useless unless one has a decryption key or password. This was critical for protecting data both in the local database on the laptop and on the server.

5.3. Use of enumerator map application during CAPI data collection

The IST platform also used QGIS, a free and open-source geographic information system software, to display addresses on a map. The QGIS software was adapted to make it more user-friendly and integrated into the IST platform for easier use by enumerators. Python code behind QGIS (PyQGIS) was used to automate tasks and create custom plugins within the software. The QGIS Python API provided access to a range of functions and tools, enabling users to manipulate data, perform analysis and create custom user interfaces within the software. With just one click, enumerators could open the map, which contained addresses assigned to them. The map was directly connected to the enumerator's address book and any changes in the address book were reflected on the map, and *vice versa*.

5.4. Data transmission system – quality and security

To maintain data consistency and accuracy, it was necessary to synchronise data between a local database and a server. This was done through a web service that enabled secure and efficient data transfer. The synchronisation process, through regular data updates, made the latest data always available. Enumerators initiated two-way data synchronisation manually as many times as necessary, although twice a day (morning and evening) synchronisation was recommended. This allowed close monitoring of fieldwork and timely responses to any challenges. Synchronisation operations were treated as a single unit of work, like transactions, to maintain data integrity and consistency. Partial updates were prevented, and all operations were either completed successfully or not at all. Internet connectivity was required for the synchronisation, and data had to be complete and marked as 'ready to send' before it could be transmitted to the server.

5.5. Data centre

The SORS used centralised server infrastructure in one data center for the census. Two virtual servers with 64 GB RAM, Intel Processor 2.8 GHz, 6 TB HDD and a robust database system were installed in the data center. High availability disaster recovery was provided by replicating data in real time between the two servers. Transactional replication in real time was used. Primary data server was dedicated to transactional processing, while the secondary data server was used for analytical purposes. This solution guaranteed that the performance of the census would be unhindered, and opened the possibility for all the necessary analyses and reporting.

Database and log backup was continuously taken on both SQL servers and there was no possibility of data loss. To ensure the highest possible level of data security and protection, census data were encrypted.

Hardware metrics (including the use of CPU and the usage of memory and disk space) were monitored to assess the overall server performance and detect any performance or resource issues.

5.6. Accuracy and security

To ensure the accuracy and security of the database, log tables (on servers and laptops) were used to track all the changes made, including data inputted by enumerators during daily data entry. These tables were used to monitor enumerators' work, identify errors, and make necessary corrections. Log tables also served as an essential tool for database security, as they allowed the tracking of unauthorised access or failed login attempts. In short, log tables were an effective means of enhancing the census performance, protecting the database from malicious actors, and ensuring the accuracy and integrity of collected data. By means of these tables, administrators could efficiently monitor user activity, identify potential threats, and take appropriate measures to safeguard the database. Overall, the implementation of log tables was crucial for achieving the goals of the survey and maximising data quality.

5.7. Web data collection (CAWI)

To address the challenge of conducting surveys in person, three web applications were developed. One of them was specifically designed for enumerating people residing in social institutions, another was intended for those living in correctional institutions, and the third one was dedicated to individuals working in diplomatic or consular outposts. These applications were developed using the aspx.net framework, and the data collected from them were stored on a separate server for later analysis.

Furthermore, various user-friendly software applications, such as those for equipment distribution and enumerator selection, were also developed to ensure the smooth performance of the census. However, for the purpose of this paper, we focused on the main applications that were described above.

6. Central Operation Control System (monitoring)

To facilitate fieldwork monitoring, a role-based web application was developed using the aspx.net framework. It provided various levels of access and permissions to different types of users based on their roles and responsibilities within the population census organisational system. The application had four different interfaces for four different groups of users: instructors, municipal coordinators, regional coordinators and supervisors. Upon login, the application provided different options depending on the user's name and password.

The primary purpose of the application was to enable real-time monitoring of the fieldwork of enumerators, instructors and municipal coordinators, through tabular views and summary reports, based on the role of the logged-in user. The application also allowed the redistribution of materials from one enumerator to another in the event of an enumerator's resignation or inability to complete their work on time. Instructors could return incorrect or incomplete data to the enumerator for revision or approve a request for data revision. Through the application, users could mark which part of the material they reviewed and access data visualisation platform reports.

An essential feature of the application was generating final worklists for enumerators and instructors based on their performance. The application had a direct connection to the server and the tables that enumerators filled by synchronising data, ensuring the accuracy and consistency of the data recorded.

6.1. Census monitoring tool

To enhance the monitoring and control of the population census process, the SORS implemented data visualisation platform for business analytics. Interactive dashboards and reports were integrated into a web-monitoring application used by the census staff responsible for the project supervision and control.

The main objective of using the data visualisation platform component was to monitor enumerators' work in near real-time and identify any anomalies promptly, to be able to make timely decisions and take appropriate measures wherever necessary. Enumerators synchronised data twice a day, and the collected information was displayed in near real-time using dashboards. Different reports were created and customised for supervisors, regional coordinators, municipal

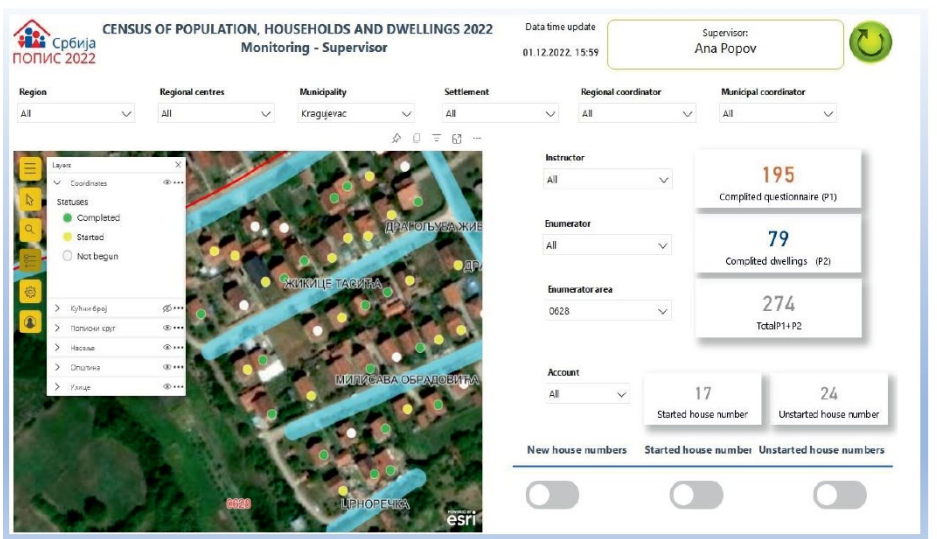
coordinators, municipal instructors, municipal census commissions and info-centres. In data visualisation platform, as in the monitoring web application, row-level security (RLS) approach was adopted to provide each user with access to only this part of work he/she was responsible for, while municipal and regional coordinators had insight into the work of their instructors and enumerators. The reports varied in size and type of data displayed depending on the user's role.

Each report provided comprehensive information on monitoring and controlling the work of all participants involved in the census. One of the key features of the reports was a map showing clear boundaries of census districts and the buildings that were surveyed.

These objects could have three different statuses (indicators):

- *not started* – all objects that were the subject of the census and were on the enumerator's list. Such objects were marked with a grey circle. At the very beginning of the census, only grey indicators could be seen on the map;
- *started* – objects in which the enumerator found a certain number of people but did not find them all. The yellow circle on the map indicated such dwellings;
- *completed* – all persons inside that building were enumerated. A green circle marked such buildings (Figure).

Figure. Example of monitoring Power BI report page on the 2022 Census of Population, Households and Dwellings



Source: authors' work.

It was possible to filter the desirable data within each report. By selecting a specific census participant, the report filtered information related only to her/him.

All important information, such as full name, telephone number and e-mail address of each participant assigned to the current report user could be seen, which greatly facilitated the communication process.

In addition, the SORS created a special Power BI report, intended for municipal census commissions. During fieldwork, the municipal census commissions were faced with large fluctuations of data. A large number of participants had changed their original status, so it was necessary to provide the commissions with a direct view of the central database, in real time. Authentication at the entry of the report ensured that each commission could see only the data for which it was responsible.

Customised Power BI report pages were created exclusively for top-level supervisors to facilitate their strategic and operational decision-making. These pages provided information necessary to determine whether it was necessary to recruit more enumerators and to identify critical territories where enumeration might have been carried out beyond the given timeframe.

Table 1. Data visualisation platform reports content on the 2022 Census of Population, Households and Dwellings

Row-level security	Specific report content	All reports content
Supervisors	Detailed side-by-side overview, and comparison of the present census with the one conducted in 2011. Insights to determine whether it was necessary to recruit more enumerators and to identify critical territories where enumeration might not be completed within the given timeframe	• Full name and role of the logged user • Time of the last data update • List of census participants and districts • Information about the enumerated population, households and dwellings • Number of completed questionnaires and rejections • Table with detailed information about the enumerated dwellings and population • Map with clearly defined boundaries of the census district and objects
Regional coordinators	The list of additional and on-duty enumerators, a side-by-side overview and comparison of the present census with the one conducted in 2011	
Municipal coordinators	Instructors and their enumerators' performance by day. Page with information which house numbers are assigned to a specific enumerator	
Municipal instructors	Enumerators' performance by day, their activity status (active/inactive), as well as the status of census districts, whether they are active or idle	
Municipal census commissions	Direct view of all participants' statuses and census staff statuses in a central database, in real time	
Info-centres	Contact information to enumerators to whom a certain address was assigned	

Source: authors' work.

6.2. Data visualisation platform metrics

The timely completion of the population census in Serbia was facilitated by the implementation of two key decisions, namely the recruitment of additional

enumerators and the reorganisation of existing enumerators based on their expertise. The identification of potential issues using the census monitoring tool, and the prompt response of the authorities to the changing situation were crucial.

Table 2. Data visualisation platform – number of report views

Power BI report	Views	Unique viewers
Info-centres	580	5
Municipal instructors	394 435	1 998
Municipal coordinators	46 031	255
Municipal census commissions	18 296	158
Regional coordinators	3 375	24
Supervisors	2 912	15

Source: authors' work.

The data collected from this metrics indicates that approximately 2,500 different viewers viewed the reports nearly half a million times. Notably, the viewers accessed the reports at least five times a day, which suggests a significant demand for the information provided.

6.3. The monitoring system and planning for enumerator reserves as key measures in overcoming population census challenges in Belgrade

At a certain point, the population census in Belgrade was facing a significant challenge. The working dynamics of the enumeration process were not aligned with the objectives and the census activities were in danger of not being completed on time. It was a critical situation that required immediate action.

With the help of the Power BI reports, the authorities were able to gain insight into the situation. They realised there was a shortage of enumerators, which was a major hindrance to the progress of the census.

As the successful completion of the census was a priority, it was necessary to begin with contingency planning for enumerator reserves. This meant developing a comprehensive plan for the deployment and management of enumerators, including training, scheduling and logistics. The plan also needed a special focus on Belgrade. In this regard, two strategic decisions were made:

- the first one was to train additional enumerators specially for Belgrade. This was expected to help address the shortage of enumerators and increase the efficiency of the census process. Additional Belgrade training sessions followed a special, fast schedule, spanning over three days (starting on the 7th, 10th, and 14th day of the census). In total, 280 enumerators were trained (a hundred enumerators trained in each of the first two sessions and 80 in the last one). These staff

memberts successfully enumerated 87,000 persons in Belgrade, i.e. 5.2% of the total enumerated population of Belgrade;
- the second decision was to regroup the best enumerators from cities in Vojvodina and Central Serbia and send them to Belgrade. However, this was done after they had completed their original tasks in the respective areas. A total of 256 enumerators from 68 different municipalities were re-hired for Belgrade, and they managed to enumerate approximately 33,000 persons, i.e. 2.1% of the total enumerated population of Belgrade.

6.4. Help Desk Ticketing System

The SORS set up an IT support team to assist instructors and enumerators during the census implementation process. The team consisted of three levels of support: IT assistants (IT1), IT staff members at the SORS regional centres (IT2), and senior IT staff members at the SORS headquarters in Belgrade (IT3). Weekly online meetings were held to improve the process. The Population Census Help Desk Ticketing System was created to provide automated communication and store solutions to hardware and software problems.

Table 3. IT support team activities during the 2022 Census

IT support level	Activities
IT1	• Receiving phone calls about specific hardware or software problems in the field • Creating new tickets • Solving problems or forwarding them to a higher level of support • Sending automated notifications to a higher level, if needed
IT2	• Solving problems forwarded by IT1 and changing ticket status to resolved • Forwarding complex problems to a higher level of support • Sending automated notifications to a lower or higher level, if needed
IT3	• Solving problems forwarded by IT2 and changing ticket status to resolved • Sending automated notifications to a lower level, if needed

Source: authors' work.

The IT support team played a crucial role in the successful census implementation by using an automated communication system. This system enabled support in order to resolve tickets at the appropriate level or forward them to the next support level for further assistance. The process was streamlined with email notifications sent automatically to higher support levels when a ticket was submitted and to lower levels when the issue was resolved.

7. Use of machine learning in post-enumeration phase for occupation and economic activity classification

In the post-enumeration phase of a census, classifying occupations and economic activities, based on the International Standard Classification of Occupations (ISCO) and the Statistical Classification of Economic Activities in the European Community (NACE), usually involves large workload (and thus requires considerable human resources). Automatic classification through data entry applications is often not feasible due to the lack of unique descriptions for different professions and activities. To reduce the time required for the classification, machine learning (ML) has emerged as a natural solution (United Nations Economic Commission for Europe Statistics Wikis, n.d.). The task involves creating an algorithm that can classify activities and occupations on the basis of a written description the interviewer inputs listening to answers to open-ended questions.

ML classification relies on large datasets of descriptions and classifications of occupations, which consider the impact of various factors such as the education level, age and gender. Although occupation and activity descriptions are a primary factor in classification, the education level has the most significant influence on it. For instance, an economist with a high school education is likely to be classified differently from the one with a university degree, even if their occupation descriptions are the same. Gender also plays a role in determining professions and activities, as some may be predominantly associated with one gender, such as miners being mostly men. Age can also carry information, as more 'modern' occupations are more likely to be associated with younger respondents, while older respondents are more likely to have traditional occupations.

To build an ML model, data sets were created on the basis of information derived from the census and other statistical surveys, including various descriptions of occupations, activities, education levels, and age and gender of respondents.

7.1. Cloud (hardware) and data anonymising

The size of the dataset generated by the algorithms is extensive, requiring more powerful hardware than a standard PC to process the data efficiently and reduce model training time.

The SORS used a cloud for ML due to its scalability, security, and user-friendly interface. Anonymisation was done before moving the data to the cloud to protect sensitive information and secure privacy. Only necessary variables were extracted from the database and a unique row identification was created to later connect the original database with the rows extracted for data processing on the cloud. After the classification using ML, data were pulled out from the cloud and transferred to the

SORS data centre to be connected with the corresponding records via unique row identification. Appropriate security and access controls were implemented (including encryption, firewalls, and IAM (Identity Access Management) tools) to prevent unauthorised access or disclosure of sensitive data.

The system used on the cloud for the ML purposes was equipped with 64 CPUs, 256 GB of RAM, an Intel(R) Xeon(R) Platinum 8370 C CPU @ 2.80 GHz 2.79 GHz processor, and a 64-bit operating system, x64-based processor.

7.2. Data set for the ML model

The initial data set for the ML model was sourced from the previous census conducted in 2011, which comprised approximately 2.5 million descriptions of various occupations and activities. However, since this data was collected using the OCR technique (the technology that allows software to interpret text on scanned images), the text obtained from scanned questionnaires was mostly unusable for the model creation. Thus, another ML model was developed to identify relevant words that corresponded to the descriptions of occupations and activities. To create the ML model, a combination of two datasets from different sources was used. The first source was the Labour Force Survey (LFS), which collected data from approximately 30,000 respondents every three months. This survey provided a constant influx of new data and included new occupations and activities not represented in the 2011 census. The second source was the Central Register of Mandatory Social Insurance (CROSO), which contained about 2.5 million records of officially registered occupations and activities for all the citizens of the Republic of Serbia who had social insurance. The use of these two data sets provided an ideal training set for the ML model, as the descriptions of occupations and activities were entered through an application, eliminating errors resulting from the manual entry. Fields chosen for the data set layout were: AGE_LEVEL, SEX, ISCO_TXT, NACE_TXT, ISCO4D and NACE3D. The textual components of the data set were transformed into a vector form using the TF-IDF (term frequency – inverse document frequency) algorithm, which generates a vector for each word based on its frequency in the total dictionary. Due to hardware limitations, TF-IDF vectoriser settings remained at: min_df=5 (remove words that are mentioned in less than 5 descriptions), max_df=0.9 and use_idf=True. With max_df=0.9, we removed from the dictionary the words whose frequency of occurrence is greater than 90%. The rationale behind this is that words which occur very often, such as ‘and’, ‘the’, and ‘is’, do not contain much information about the content of a data set. By using max_df, we excluded such terms from our feature set.

The resulting vector contains approximately 13,000 members. However, the matrix size produced from the vector is enormous and requires substantial hardware resources. To address this challenge, dense matrices were converted to sparse matrices.

7.3. ML algorithms used during training

Two approaches, i.e. classic classification algorithms and neural networks were considered when decisions about the classification method were made. While the latter are known to offer higher accuracy, their computational demands and requirement for extensive fine-tuning were deemed impractical given the available timeframe. As such, the SORS opted for using the common ML algorithms instead.

Three algorithms were tested for accuracy: logistic regression, random forest classifier, and XGBoost. After initial testing, the random forest classifier was chosen due to its superior accuracy compared to the other two algorithms (61.08%).

The hyperparameters used during GridSearchCV with validation across three different mixtures of the training dataset are: {'n_estimators': [50, 60, 80, 100, 150, 160, 200], 'max_features': ['log2', 'sqrt'], 'max_depth': [80, 90, 100, 110, None], 'min_samples_split': [2, 5, 10, 20, 50], 'min_samples_leaf': [1, 2, 4], 'bootstrap' : [True, False]}.

The best hyperparameters obtained for the best model are: {'n_estimators': [160], 'max_features': ['log2'], 'max_depth': [None], 'min_samples_split': [2], 'min_samples_leaf': [1], 'bootstrap': [True]}.

The classification accuracy achieved 98% after training and validating on the training set. Classification accuracy testing was conducted on a sample checked by coders, and the reports were satisfactory.

The SORS's next step was to train the ML algorithm to correct the words from the previous census, thereby enriching the data set with 2.5 million new entries.

8. Census undercoverage and use of administrative data for item imputation

Given the purpose of the population census – to provide a comprehensive picture of the entire population, including citizens who did not participate in the census for whatever reason (out of country, did not want to participate, not found on the address), the SORS's efforts to produce high-quality census data do not end when the fieldwork is done. They continue during the data processing stage.

It turned out during fieldwork that the response rate in some urban areas was lower than in previous censuses. However, the fact that the SORS had access to all relevant administrative registers, including the Personal Identification Number

(PIN), made it possible to use a combined census method. The combined census in this case means that a database obtained through the direct census interviews is supplemented by administrative data for those citizens who did not participate in the census for whatever reason.

Most census topics are not fully covered by one administrative data source, but are likely to be partially represented in several different sources. Different approaches and methods are necessary to combine information and draw full information from a range of sources. In line with the census methodology, the administrative sources used to determine the permanent resident population in Serbia are: the recently established Central Population Register as well as administrative registers on the employed and unemployed, students and children attending kindergartens, retired people, taxpayers, beneficiaries of social assistance, beneficiaries of health insurance, etc. All the available administrative data were integrated into one master administrative dataset. Because the Central Population Register contains data on all citizens of Serbia, regardless of whether they live in the country or not, additional information from other listed records (called 'signs of life') was used to determine the groups of the population that definitely lived in Serbia on the Census Day, and as such corresponded with the definition of permanent residents. The record linkage of census and administrative data was performed by means of both the deterministic and probabilistic method.

Before matching the census and administrative data, preprocessing tasks were carried out to maximise the quality of the data linkage. These activities included:

- cleaning the matching variables (e.g. name, date of birth, address), which means that the variables were converted to an adequate format; null strings were ignored; abbreviations, punctuation marks, upper/lower cases, etc. were cleaned; and all the necessary transformations were carried out to standardise the variables;
- deriving new matching variables (for instance, month and year of birth);
- removing some records from dataset: (a) duplicates missed by the resolve multiple responses process, (b) individuals born after the Census Day, (c) records missing names and date of birth.

The second task was to choose variables for the record linkage. The parameters used for matching records from the census and administrative registers were PIN (where it was available and correct), or a combination of the following parameters: the municipality of birth, the date of birth, name, surname, parent's name, the municipality of residence and home address (in the cases where the PIN was missing or inaccurate). The high degree of certainty required for the linkage was achieved because of the presence of PIN in both datasets. About 80% of individuals in the census provided the correct PIN.

To improve the scalability of the linkage process, the creation of blocks was carried out using the following variables: name, surname and the date of birth, to limit the comparison of records to pairs or groups that likely correspond to the same entity. Subsequently, deterministic record linkage was applied to several subsets of the aforementioned parameters. As a result of this phase, many records were successfully matched, and the working datasets were significantly reduced.

Afterwards, some further analysis and probabilistic record linkage were done on the restricted datasets without records that were already matched. With smaller unmatched datasets, both from the census and administrative data, it was possible to perform some more detailed and time-consuming methods. Two of the algorithms that were used most frequently for comparing strings were the Soundex and the Levenshtein distance algorithm.

In the last stage, the matching of records between the census and administrative master datasets required significant clerical effort to find matches that could not be found by automatic means.

The process described above yielded two datasets: the first (A dataset), used for the imputation of the census data – individuals existing in administrative registers and considered as usual population but unaccounted for by the census (≈218,000) and the other (B dataset), consisting of individuals not found in the administrative data but present in the census dataset (≈50,000). In the A dataset, individuals were grouped in households wherever it was possible according to the address or common children living at the same address. Then associative clerical matching was used to check matched households containing unmatched persons. Unmatched households containing matched persons were also investigated to determine whether the households should also be matched. It was possible to make new person matches, or to break the existing person and household matches. This enabled the optimal within-household person linkage. Clerical work was also performed to further explore the dataset of individuals present in the census but not found in administrative registers.

9. Statistical Population Register

Over the past few decades, administrative registers have grown in importance as a source of data for official statistics. They offer numerous advantages, including cost reduction, improved data quality, and the ability to produce more frequent information.

Following trends prevailing among the EU member states, the transition from a traditional approach to the one relying on the use of administrative data for statistical purposes is a part of the SORS's vision of developing a more general

register-based statistical system. To fulfill this vision, bearing in mind that transition is a long-term and complex task, the SORS continues its work on the creation of conditions that would make it possible to conduct the next census (in 2031) on the basis of administrative data.

On the way to the introduction of a census based on register data, numerous activities have been planned. The first step after the completion of the 2022 Census is carrying out an exercise to simulate a 'register-based census' using administrative sources available at that time point, and to compare the administrative data with data obtained through the census.[2] As a result, the quality of both main registers and supplementary registers will be assessed, showing the potential shortcomings in terms of availability, coverage, concepts, reference date, accuracy, etc.

In this regard, it is planned to implement the following:

- the SORS is going to prepare a road map with the proposed targets and to define necessary steps to be taken on the basis of the main findings from the previous activity;
- afterwards, national workshops will be organised to bring together the representatives of relevant institutions to consider the possibilities, future steps, deadlines, roles and responsibilities in the process of introducing a register-based census;
- the conclusions arising from the workshops will be presented by the SORS at a conference;
- the final step will involve drawing up the action plan for the introduction of a register-based census and identifying all public administration bodies relevant for this activity.

Considering that it is not possible to collect all the necessary census variables (core topics) defined by the Recommendations (United Nations, 2015, 2017) from administrative sources, the SORS, relying on examples of good international practice, decided to overcome the lack of certain data by establishing a statistical population register (SPR).[3] The need for the SPR was already recognised in the previous census cycle, but creating it then was not feasible due to the absence of essential administrative registers like the Central Population Register. The establishment of the SPR has became possible thanks to the improvement of the national system of administrative registers in recent years.

The SPR will be established on the basis of the 2022 Census data and will be regularly updated using information from available administrative sources. The master file of the SPR will consist of the data from the 2022 Census. All the core

[2] Within the *IPA 2018 National Programme – EU for Development of Statistics in Serbia* project.
[3] Within the *IPA 2022 National Programme – EU support for efficient statistics and further strengthening statistical infrastructure* project.

variables defined by international recommendations, which are available and of satisfactory quality, will be regularly taken from administrative registers for updating the SPR. For example, the level of education will be updated from administrative sources for persons who continue receiving education; otherwise, a person's education status will remain the same as it was stated during the census. Variables defined as non-core according to international recommendations (such as means of transport, etc.), will be available only from the 2022 Census. So administrative registers will be used as sources for the statistical register, but the reverse process will not be possible, respecting the principle of 'one-way flow'.

The SPR will be one of the key preconditions for the 2031 census to be carried out as register-based. In addition, the SPR will be an up-to-date sample frame for running statistical surveys, and it will contribute to the annual production of selected population datasets foreseen by EU legislation.

10. Conclusions

The use of digital technology in censuses has become increasingly important due to the advantages it offers over traditional paper-based methods. In the SORS's experience, the digital census improved the accuracy, reduced costs, and speeded up the process of data collection, processing, and dissemination. Additionally, it allowed the adoption of advanced technologies such as geospatial technology, machine learning, record linkage or imputation, which, in turn, allowed a more efficient and effective analysis of the census data.

Compared to a register-based census, traditional census based on a digital technology provides more detailed and accurate information on the demographic properties, such as ethno-cultural characteristics or household and family composition. Traditional census may be more inclusive by ensuring that everyone is counted, regardless of their participation in government programmes or their legal status. However, traditional censuses are more expensive, time-consuming and labor-intensive than register-based censuses.

The SORS, on the basis of its experience from 2022, believes that a digital traditional census, without self-enumeration, that adopts administrative data for item imputation, is the most effective way to obtain precise and comprehensive information about a population, including its demographic characteristics, geographic distribution and the overall size.

References

Law on 2021 Census of Population, Households and Dwellings (Official Gazette of RS No. 9/20 of 4 February 2020). https://popis2022.stat.gov.rs/media/1596/law-on-the-2021-census.pdf.

United Nations. (2015). *Conference of European Statisticians Recommendations for the 2020 Censuses of Population and Housing*. http://www.unece.org/publications/2020recomm.html.

United Nations. (2017). *Principles and Recommendations for Population and Housing Censuses, Revision 3*. https://unstats.un.org/unsd/demographic-social/Standards-and-Methods/files/Principles_and_Recommendations/Population-and-Housing-Censuses/Series_M67rev3-E.pdf.

United Nations. (2019). *Guidelines on the use of electronic data collection technologies in population and housing censuses*. https://unstats.un.org/unsd/demographic/standmeth/handbooks/data-collection-census-201901.pdf.

United Nations Economic Commission for Europe Statistic Wikis. (n.d.). *Machine Learning for Official Statistics*. https://statswiki.unece.org/display/ML/Machine+Learning+for+Official+Statistics+Home.

Zakon o izmienama zakona o popisu stanownisztwa, domaćinstawa i stanowa 2022. Godine (Official Gazette of RS No. 35/21 of 16 April 2021). https://popis2022.stat.gov.rs/media/1595/zakon-o-popisu-2021.pdf.