CONFERENCE OF EUROPEAN STATISTICIANS

Workshop on Statistical Data Collection

10-12 October 2018, Geneva, Switzerland

Session 5

24 September 2018

The Actions Taken To Enable The Administrative Records To Be Used By Subject Matter Units, The Difficulties Experienced And The Proposed Solutions In Turkstat

Bilal Kurban (Turkish Statistical Institute)

BILAL.KURBAN@tuik.gov.tr

Abstract

Due to reporting burden, difficulties in direct collection from business respondents and increasing costs, Turkish Statistical Institute (TurkStat) has decided to extend the use of administrative data in statistical production. This has led TurkStat to redesign and modernize its business statistics system on the basis of administrative records.

Under the cooperation and data exchange agreements between TurkStat and administrative authorities and for the purpose of increasing and expanding the use of administrative data in business statistics, tax records from the Revenue Administration (RA) and Social Security Institution (SSI) data which provides social security records of individuals have been shared with TurkStat. This has been a milestone for TurkStat's official statistics on business and economy. Thus, using administrative data directly or indirectly in the production of indicators in especially short term and annual business statistics and national accounts have been started.

On the other hand; since administrative records are not collected for statistical purposes but for the institutions and organizations to carry out their own business and transactions, their concepts, definitions, reference dates and scope may differ. For this reason they cannot be used directly in the statistical production. They first need to be harmonized and linked with the statistical business registers for the purpose of anonymization of data for internal use and code assignments. Moreover, technical breakdowns and possible problems may occur when transferring administrative records, which requires some basic validation of the structure and integrity of the information received before using them for statistical production.

The quality of each processing phase up to the realization of the statistical product has its own importance, but the quality of the collection phase of administrative information also significantly guarantees the quality of other phases of the process (Dhuli, 2018). In this context, from the perspective of the RA and SSI records, the paper tries to describe what steps the administrative

records go through before subject matter units use them, difficulties experienced in the process and the solutions developed to overcome them in TurkStat.

References

Elsa Dhuli, (2018), Implementation of General Statistical Business Process Model in the administrative source.

The Actions Taken To Enable The Administrative Records To Be Used By Subject Matter Units, The Difficulties Experienced And The Proposed Solutions In TurkStat

Bilal Kurban[1], Hatice Burcu Eskici[1], Mahmut Öztürk[1], Serkan Arslanoğlu[1] and Muhammed Fatih Tüzen[2]

[1]TurkStat Expert
[2] TurkStat Assistant Expert

1. Introduction

Due to reasons such as reporting burden, difficulties in direct collection from business respondents and increasing costs together with widespread changes in the social and technological context have required Turkish Statistical Institute (TurkStat) to improve the quality of data collection phase of statistical production process while reducing reporting burden. This has led TurkStat to redesign and modernize its business statistics system on the basis of administrative records.

Under the cooperation and data exchange agreements between TurkStat and administrative authorities, records from Revenue Administration (RA) which keeps tax records and Social Security Institution (SSI) have been shared with TurkStat, which has been a milestone for TurkStat's official statistics on business and economy.

The quality of each processing phase up to the realization of the statistical product has its own importance, but the quality of the collection phase of administrative information also significantly guarantees the quality of other phases of the process (Dhuli, 2018). In this context, from the perspective of the RA and SSI records, the paper tries to describe what steps the administrative records go through before subject matter units use them, difficulties experienced in the process and the solutions developed to overcome them in TurkStat.

2. Reasons for Change in Business Statistics

Among the motives of TurkStat to redesign and modernize business statistics based on administrative data are:

- Reporting burden,
- Difficulties in direct collection from respondents,
- Increasing costs,
- Developments in technology,
- Constantly growing data size and increasing demand for data,
- Difficulties in measuring economy through primary data sources,
- Measurement errors and
- Good practices in other national statistical offices.

Due to these reasons, in order to be able to produce data in a way that it meets national and international needs, TurkStat has begun to benefit more from the use of administrative records in business statistics. For the production of business statistics, the most common administrative data source worldwide is tax returns of businesses. Records kept by social security institutions play an important role, as well (Eurostat, 1999).

3. Redesign of Business Statistics Production Based on Administrative Records

Thanks to statistical law of Turkey that It allows access for TurkStat to administrative data. Through data exchange agreements between TurkStat and administrative authorities, TurkStat has begun on-demand and scheduled administrative data transfer from RA and SSI for the purpose of business statistics production since the second half of 2017. Afterwards it has been possible to acquire the micro level data of these two institutions. Subsequently, metadata (definitions, scope, etc.) of the variables included in the administrative records were studied and the variables to be used in the statistical production were determined. This process has taken time due to the data size, need for high-level technical knowledge to understand administrative characteristics and the depth of the legislation, need for linking the data and analyzing the consistency.

With the integration of new administrative data sources into Turkish Statistical System, domains in TurkStat affected most by the process of extending the use of administrative data in business statistics have been business registers, short-term and annual business statistics and national accounts. Starting from March 2018, using administrative data directly or indirectly in the production of indicators in especially short term and annual business statistics and national accounts have been started.

4. Difficulties Experienced in the Integration Process

The first difficulty that had to be overcome to integrate new and big administrative sources into the system was the technical and methodological uncertainties associated with the administrative sources. At first, exactly what and how to transfer and to what extent it will be consistent with the expectations was initially uncertain.

Data storage, where and how to store the terabytes of data containing more than 1500 tables, was another challenge. Transferring all the tax and social security data means that TurkStat met with a volume of data that had never been exposed before.

Next, as the transfer data are not produced for statistical purposes but only for administrative record owners to fulfill their functions there was limited data about the data other than the legislation itself.

In addition, as the source of data collection shifts from primary to secondary, TurkStat needed to have a new employee profile having combined capabilities of IT and statistics, which meant the beginning of a learning and training process in an unprecedented way in TurkStat.

Lastly; having the best strategy is no use if the institutional culture is not open to change. Therefore in statistical organizations, it should be created a new mindset embracing the change, so should in TurkStat.

5. The Operations Performed on Administrative Data before the Use of Subject Matter Units

Since administrative records are not collected for statistical purposes but for the institutions and organizations to carry out their own business and transactions, their concepts, definitions, reference dates and scope may differ. For this reason they cannot be used directly in the statistical production. They first need to be harmonized and linked with the statistical business registers for the purpose of anonymization of data for internal use and code assignments. Moreover, technical breakdowns and

possible problems may occur when transferring administrative records, which require some basic validation of structure and integrity of the received data before using them for statistical production.

The operations performed on administrative data before the use of subject matter units are listed below respectively.

5.1. Reminder Sending

The process starts with sending reminder. An e-mail reminder is sent to administrative record owner one day before the agreed transfer date to remind the planned data transfer. Transfer dates are specified as annually with source data experts before the New Year begins.

5.2. Transfer of Raw Administrative Data to TurkStat

This step includes checking if the files are transferred to the specified temporary server by the administrative source and if yes, transferring the files to a workstation. ETL tool is used to transfer data in cases where the database used in TurkStat differs from the administrative record holder.

5.3. Initial Controls of Transferred Data

By comparing the names of the incoming tables and the expected tables, it is checked whether there are missing tables. If there are any missing tables, they are requested again from administrative record holder. In case there is no missing table, this time, it is checked whether there is data in the tables (empty table control). In case of empty table(s), administrative data owner is asked to resend the empty table(s).

5.4. Transfer of administrative data from the workstation to related database scheme

After the initial checks, administrative data tables are transferred from the workstation to the related database schema. There are different schemes in the database for each administrative data set. At this stage it is verified that there is no problem in transferring all the files smoothly.

5.5. Creating Indexes for the Tables

In order to allow queries to efficiently retrieve data from the database, indexes are created for the tables in the database. The indexed tables are now ready for structure and integrity checks.

5.6. Structure and integrity checks

Structure and integrity checks which are crucial for timely detection and fixing technical breakdowns and possible problems that may arise in transferring administrative records to TurkStat are made according to the data validation rules developed by TurkStat.

Structure and integrity checks also include checking historical changes of administrative data variables and examining these changes.

Structure and integrity checks should include at least the following:

5.6.1. File sizes

Each file included in the transfer plan has an expected size. It is expected that the size of the incoming files should be close to the expected size (and greater than 0). Table sizes are one of the most important checks in early detection of incomplete and/or excessive (usually due to duplication) data transfer.

5.6.2. Number of tables and table fields

The total number of tables transferred from the administrative source and table fields is checked to see if it is the same with the agreed number of tables and fields. The smaller number of tables and/or fields than expected is the other sign of incomplete data transfer.

5.6.3. Table and field names, types and empty fields

After clarification of the table and field numbers included in the transfer data, it is checked whether these tables and fields are the expected ones, whether the table and field names are correct, whether there is a difference in the data types and whether there are empty fields. The diagnosis of the differences encountered during previous numerical controls is made here. In fact, there may still be differences between the expected and the observed, even if there is no problem in numerical comparisons.

It is expected that the table and field names and types will be the same as the previous transfers or agreed structure (if exists). If there is a difference (such as adding a new column, removing the existing column or changing the name of the existing column) these differences are clarified by consulting with the authorities of the administrative register. As the result of consultation, if the changes are persistent, the necessary changes are made for the next transfers. In addition, if the type of table fields has changed, this may have caused data loss.

Moreover, table and field names and types are critical in the later execution of written automated procedures. Differences in the names and types may cause the procedures not to work properly. Once names and types are not checked properly at the beginning, later it is much more difficult to trace the procedure and reveal the source of the problem.

In this step, it is also checked whether there are empty table fields. It is expected that the fields (with the exception of deliberately empty ones) are not empty.

5.6.4. Consistency of row counts of tables in overlapping periods

If the transferred data is cumulative, this means there are overlapping periods between transfers. In this case consistency of row counts of tables in overlapping periods should be checked. The formation of administrative data is subject to legislations therefore records may be added or deleted or updated retroactively, this is taken into account when checking number of records in overlapping periods.

In the cumulative data acquisition the data received in the previous periods will be retrieved again. Therefore, in the last transfer, it should be carefully checked that there is no loss also in the data of previous periods for revisions. Due to this risk, the validated previous transfer data is also backed up in the database.

5.6.5. Expected row counts of tables for reference period

Row counts of each table are checked both whether the number of records in the tables for the reference period is between the upper and lower control limits created by statistical methods based on historical data series and whether incoming data have a radical deviation from the historical data series. If the row number of any table is not within the confidence interval or there is a high deviation from the historical data series, this may indicate that data is transferred deficiently and may also mean that there is a radical change in legislation that we are not aware of or data owner has made a change to the scope of the transfer data without notifying us. In such a case, the source of the problem is revealed in cooperation with the administrative registration authority. Once the problem has been posed, if there is a defect in the transfer then the transfer can be restored and the problem is resolved. But if the problem is due to legislation change or other reasons then more sophisticated solutions are needed.

5.6.6. Duplicate records

If the database in the administrative source was not designed to check for duplicates or if the other party sends you records using data matching techniques rather than as it is held in the database duplication problems may occur. Duplicate data can be either multiple tables containing the same data or records of the same table containing for some (or all) fields with similar data. Whatever the cause for duplicate data, it is required to remove (or to flag) existing duplicate records within the database to achieve high quality data in the end.

5.6.7. Presence of high value-added statistical units

The availability of business units contributing most to the statistical production in the transfer data is checked for the reference period. This may indicate an incomplete data transfer or another problem (e.g. unit non response).

5.6.8. Descriptive statistics of basic variables

In this step, descriptive statistics are obtained regarding the distribution of the key variables to be used in the production of statistics. Descriptive statistics detect potential problems in the variables caused by digit errors during transfers while they provide information on outliers and variability.

While the control of the transfer data is carried out in accordance with the above sequence, if there is a problem in any of the sequential steps this must be clarified before proceeding. If necessary, the data transfer must be requested again partially or fully.

As a result of basic validations if there is a problem with the data then fully or partially re-transfer of troubleshot data may require. If there is no problem in the tables to prevent validation as the result of the structure and integrity checks, administrative data is then matched and linked with the statistical business registers.

5.7. Anonymization, Classification and Coding

Following the structure and integrity checks on administrative data, before enabling subject matter units to use the data, they are matched with the statistical business registers. Business registers are

the center of the integration for business statistics and they provide the link between the administrative input data and the statistical output data.

Matching with business registers is done by using identification data fields such as tax identification number and citizenship number. Once basic validations of the data and linking have been completed, assignment of classification codes is performed at this step. In case of determining the activity code of a new born enterprises, the statistical units that meet criteria such as economically activeness, economic size (by turnover, number of employees) are sent to TurkStat regional offices by their respective priority for identifying on site their activity codes. Main activity codes are then assigned to the units checked by regional office staff on site.

In addition, data anonymization which is necessary for in-house use of individual data is also performed in this step. Anonymization of personal data is achieved by assigning private anonymous identifiers for internal use instead of personal registrars originating from the source such as administrative registration data, business registers, and address based population registers.

And finally reference tables including connection variables are created in order to enable subject matter units of TurkStat to be able to use it.

After the anonymization, classification and coding step, tables are made available to the subject matter units by creation of table views (virtual tables) and authorization of the related units. Views do not require any physical space in the database and are good way to present data in particular users from accessing the table directly.

Subject matter units now can run their procedures and carry out further logical controls and internal checks in order to control the internal consistency of the data after the integration of administrative data from different sources. As a result of the controls; miscoding, potential problems in the data, errors, outliers, item and unit non-response, unreliable data are determined and such data are edited or imputed (if necessary) using statistical methods.

Only after these steps have been successfully accomplished can statistics be calculated and published.

Management of all the above process steps is carried out via a platform which is a free and open source, web-based project management and issue tracking tool. It allows users to manage multiple projects and associated subprojects. With this platform, units and persons responsible for each process step and followers are defined in advance so that who will do and what to do is definite at the very beginning and all related followers can monitor the steps transparently. Work steps trigger each other; the previous work step is the trigger of the next one by sending automated e-mail to the corporate e-mail addresses of responsible persons, their managers and followers when there is a change or progress in any process step.

Conclusion and Remarks

The quality of each processing phase up to the realization of the statistical product has its importance, but the quality of the collection phase of administrative information and the implementation of statistical models in their holders also significantly guarantees the quality of other phases of the process (Dhuli, 2018). In this context, from the perspective of the RA and SSI records,

the paper tried to describe what steps the administrative records go through before subject matter units use them, difficulties experienced in the process and the solutions developed to overcome them in TurkStat.

In TurkStat's current practice, the data which had been previously collected from enterprises by surveys are now partly or completely compiled from administrative records of RA and SSI. Business registers plays an important role in extending the use of administrative data in business statistics. Data integration is achieved based on statistical business registers so that administrative files get together. Consequently, the coherence between the data sets of annual and short-term business statistics and national accounts has been ensured along with maintaining a decrease in response burden on enterprises, saving of time and labor and increase in data quality, as well.

The biggest problems TurkStat faced when integrating RA and SSI records were uncertainties, lack of metadata, data storage and lack of skilled data analysts to make use of large volume of data. Working in cooperation and collaboration with source data experts as well as removing technical and methodological uncertainties, metadata documents were also created. On the other hand, since hiring new employees specialized in large volume data is not always the case for various reasons TurkStat have re-skilled its current workers and database performance tuning and query optimization have became an important issue such as working with table views, indexing/partitioning etc. As data gets larger and more complex, data storing and processing habits of the staff changed. This allowed them to develop new skills and a new mindset with the help of trainings and learning by doing. Anything it had to be done repeatedly has automated, which took longer the first time but the other times time has been saved and subject matter experts has focused on the analysis. Institution-wide standards on verifying all new administrative data before it enters to the statistical system has been set and it has been ensured employees are aware of these standards. Moreover, shifting data collection in business statistics from primary to secondary has contributed to development of organizational culture positively.

The increasing use of administrative data in production of official statistics, led TurkStat to adopt new approaches regarding compilation, processing and dissemination processes of data. These new approaches, which require close cooperation between TurkStat and other data provider organizations will extend the quality of official statistics, improve analysis skills in the related institutions and contribute positively to the quality of registers, themselves.

The existing administrative registers in other data provider institutions are established for the purposes other than statistical production. Therefore, some differences occur in data regarding coverage, classifications, reference dates etc. In the meantime, Turkstat continues its efforts to keep a sustainable cooperation with other data provider organizations in order to eliminate these issues by improving and upgrading administrative registers and to maintain continuity of data delivery.

The strength and weaknesses of data derived from administrative data compared to survey data impels statistical offices to determine the effective use of these data sources. Administrative data directly or indirectly replaces survey results and they are currently used for establishing frames and data analysis. Henceforth, the use of administrative data penetrates most of the statistical domains in TurkStat and it is expected that their usage will expand further in the near future.

The era of digitalization with a new data environment and ecosystem, provokes a raise in data demands from TurkStat both in terms of volume and variety. TurkStat is in an effort to generate solid coordination mechanisms and use multiple data sources (surveys, administrative data) when producing official statistics. Integrated statistical production process is only possible through upgraded analytical skills and training human resources as data scientists. It is now an obligation for TurkStat to adopt procedure-oriented and process-motivated approaches, as currently, in statistical production, the data tables measured in "megabytes" are replaced with those measured in "gigabytes". Being aware of all these, TurkStat handles its studies by continuously and carefully monitoring its international peers and investing both in human-resources and technological infrastructure.

References

Eurostat. (1999). Use of Administrative Sources for Business Statistics Purposes:Handbook on Good Practices, Eurostat, Luxembourg, ISBN: 92-828-8024-9, http://ec.europa.eu/eurostat/documents/3859598/5854781/CA-24-99-897-EN.PDF/c13e6549-ff0c-495b-b0c8-776a86d9359c?version=1.0,[Visited: 27 February 2018].

Find duplicate records with a query, https://support.office.com/en-us/article/find-duplicate-records-with-a-query-3cc805a2-2a13-4439-b0d3-6b23c7d60fbb, [Visited: 07 September 2018].

Norbert Rainer, (2009), Business Registers as a Tool for Linking Various Administrative Data Sources.

https://stackoverflow.com/questions/256700/what-is-a-view-in-oracle

https://www.redmine.org/projects/redmine/wiki

https://www.piesync.com/blog/top-5-problems-with-big-data-and-how-to-solve-them/

Elsa Dhuli, (2018), Implementation of General Statistical Business Process Model in the administrative source.