**Integration of New Administrative Data Sources into Turkish Statistical System**

Bilal Kurban, TurkStat Expert, Administrative Registers Unit, bilal.kurban@tuik.gov.tr

H. Burcu Eskici, TurkStat Expert, Administrative Registers Unit, burcu.eskici@tuik.gov.tr

H. Tancan Kale, TurkStat Expert, Head of Labor Input Indicators Unit, tancan.kale@tuik.gov.tr

M. Arda Yılmaz, TurkStat Expert, Business Registers Unit, arda.yilmaz@tuik.gov.tr

Mahmut Öztürk, TurkStat Expert, Head of Administrative Registers Unit, mahmut.ozturk@tuik.gov.tr

Serkan Arslanoğlu, TurkStat Expert, Administrative Registers Unit, serkan.arslanoglu@tuik.gov.tr

**Abstract**

*TurkStat aims to collect and process data more quickly, more accurately and more effectively than before. On the other hand; unlike the past, it seems like data collection works have been changing rapidly in recent years. Emerging technologies, respondent burden concerns, increasing costs, globalization, having more difficulties in collecting data than the old times and exponentially increasing amount of data day by day and thus increasing data demands are among the reasons for these changes.*

*Overcoming the data collection challenges, TurkStat has given priority to "increasing the weight of administrative registers in statistical production" in its future plans and programs and has accelerated the studies on this subject. TurkStat started to pay more attention to take advantage of administrative sources. In this context, TurkStat started regular monthly data transfers from Revenue Administration as of June 2017 and from Social Security Institution as of October 2017. Starting from 2018, as well as structural business statistics characteristics, short term business statistics indicators such as turnover, production and labour input are planned to be produced by using these administrative records.*

*Furthermore, as data collection shifts to secondary sources due to integration of new data sources, TurkStat started to have a new employee profile having combined capabilities of statisticians and IT specialists. Making use of new administrative data collection activities in TurkStat also started to cause developing new skills, creating a mindset and culture open to change and developing strategic partnerships with administrative registers holding public institutions .*

*This paper shows TurkStat’s experiences on integration of new administrative data sources into Turkish Statistical System, the resulting necessity of redesigning its business statistics and analytical skills and competencies developed in handling large and complex administrative datasets.*

**Keywords:** Administrative data; data integration, business statistics

**1. Introduction**

Ever-increasing demand for statistical data both increases response burden and consumes limited budgets and resources. In addition to this, 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 response burden. In order to meet this requirement, TurkStat has given priority to increasing the use of administrative registers in statistical production.

Using administrative records for statistical purposes is not new to TurkStat: administrative records have already been made use of in statistical domains such as population, demography, labor market, education, culture, sports, social protection, crime, justice, elections, tourism, international migration, energy, transportation, foreign trade, communication and finance. However, business statistics produced by TurkStat were predominantly based on the primary data collection. Nowadays, the use of administrative data has also become a necessity for TurkStat’s business statistics rather than an option.

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, records from Revenue Administration (RA) and Social Security Institution (SSI) have been shared with TurkStat. This has been a milestone for TurkStat’s official statistics on business and economy, thanks to which, 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.

This paper explains TurkStat’s motivation to integrate new data sources, namely, RA and SSI, into Turkish Statistical System, its experience along the process and redesign of its business statistics as a result.

**2. Reasons for Need for Change**

Main motivations that have led TurkStat to redesign and modernize its business statistics system on the basis of administrative records are the following:

* Response and 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.

For these reasons, in order to be able to produce data in a way that meets national and international needs, TurkStat has begun to take advantage of the use of administrative records in business statistics.

**3. New Administrative Data Sources Integrated into the System**

Thanks to statistical law of Turkey that It allows access for Turkstat to administrative data. Article 9 (Access to administrative data) of Statistics Law of Turkey states that:

“In order to be used for the production of official statistics, the institutions and organizations are obliged to submit or open to use of TurkStat the records and other data files they compile, process and store regarding their working areas, and all kinds of maps and data obtained by remote sensing within the specified period and free of charge.”

As 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). In light of this information and best practice examples of other NSIs TurkStat has begun on-demand and monthly scheduled administrative data transfer from RA and SSI for the purpose of business statistics production since the second half of 2017.

While SSI data provides social security records of individuals, RA presents tax records as the main data sources used in business statistics. Tax records used directly or indirectly in the production of business statistics are corporate and income tax returns and supplements (balance sheets, income statements etc.), provisional tax return and supplements, value added tax, special consumption tax , withholding tax returns and buying and selling statements.

**4. Management of the Integration Process of New Administrative Data**

After gaining access to the administrative data set, it had to assess the usefulness of them for potential statistical uses because the data is collected not for statistical purposes but for the institutions and organizations to carry out their own operations. Therefore; concepts, definitions, reference dates and scope needed to be harmonized to combine data from multiple administrative data sources and to apply data quality standards to each data source.

*4.1. Need Analysis, Taking Inventory and Metadata Creation*

At the beginning of the integration process; first of all, with a similar approach as in agile project management, working and focus groups were formed with the participation of related coordination and subject matter units within TurkStat and they first carried out a preliminary needs analysis. According to the needs of the units from RA and SSI records, the outline of the data to be transferred is specified.

Following the identification of needs, all tables and table fields of administrative records to be transferred were taken inventory. Table relationships needed for each field in the inventory was produced in collaboration with administrative data owners.

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. Content, definition, scope, table names, relations between tables and table field descriptions of RA and SSI administrative data and how to use it in practice were documented by the cooperation of TurkStat and administrative data holders.

*4.2. Identification of Monthly Flow Data Variables from Overall Data and Determination of Business Characteristics Produced Based on RA and SSI Records*

Since it is required to contact with the data (because it is an empirical process) overall past RA and SSI data were transferred from the sources to Turkstat servers and opened for the examination of subject matter and other related units in Turkstat. Both working with the real life data and reviewing tax and social security regulations, focus groups refined the needs iteratively and specified them to minimize monthly flow data volume. Variables that are not meeting or partially meeting the needs were also identified.

With the concrete, numerical and comparative studies of subject matter units on the real life past data, the characteristics to be produced based on the administrative records have been determined, which were previously collected by the field survey.

*4.3. New Employee Profile and Organizational Culture*

As the source of data acquisition shifts from primary to secondary, TurkStat has to have a new employee profile having combined capabilities of IT and statistics. With the integration of RA and SSI data sources, it has become even more important for TurkStat to develop internal analytical skills and competencies and to support data analytics skills among its staff which meant the beginning of a learning and training process in an unprecedented way and a new mindset embracing the change. Transferring all the RA and SSI data means that Turkstat met with a volume of data that had never been exposed before; more than 1500 tables and terabytes of data.

In order to overcome the problems caused by integrating large-scale data into the system, efforts have been made to adapt TurkStat staff to new situation and started to provide them with new skills. Process of learning and adaptation to new culture started immediately by the help of trainings and working together (collaborative works between and within units). Eventually, data processing habits of the many of the staff have changed. Previously, easy-to-use data processing tools having simple interfaces were sufficient for processing and analyzing data when table dimensions were measured in megabytes. In the new case, however; as data gets larger and more complex having tables measuring hundreds of gigabytes, it became necessity for the stuff to process data directly on the database environment by writing SQL and procedural SQL codes. And database performance tuning and query optimization techniques such as; working with table views, indexing and partitioning became one of the most key issues.

*4.4. Business Registers Based Integration*

It is not always easy to combine an administrative data source with another source of information. This is especially true if there is no common identifier for both sources as in the case of RA and SSI records. It is required the use of record linkage techniques. Business Registers (BR) which is set up based on administrative registers by TurkStat plays an important role in extending the use of administrative data in business statistics. BR plays as an intermediary between administrative data input and statistical output. In other words, BR serves as a linking system to administrative databases.

**5. Statistical Production Process from Administrative Data**

RA and SSI records are not intended to be used directly for statistical purposes and can not be used directly; quality control procedures should be carried out on them. Following the transfer of administrative data to TurkStat, a series of sequential processing steps such as validation, editing and imputation based on documented procedures and matching and linking of different data sources including administrative ones must be carried out.

*5.1. Data Acquisition*

The first phase of the statistical production process from administrative sources is the acquisition of data from the owner of administrative records. The processes relating to operations performed from the data acquisition to data processing phase is;

* data transfer from administrative source to TurkStat,
* basic validations (structure and integrity checks),
* anonymization, classification and coding and
* authorization of subject matter units for data processing ,respectively.

After transferring the administrative data to TurkStat, first, the indexes of the transferred tables are created. And then structure and integrity checks are made according to the data validation rules developed by TurkStat, which is crucial for timely detection and fixing technical breakdowns and possible problems that may arise in transferring administrative records to TurkStat.

 Basic validations include the following checks:

* File sizes
* Number of tables
* Table names and number of table fields
* Field names, types and occupancy
* Number of Rows in the overlapping periods
* Row counts of tables and whether they are within the expected upper and lower control limits
* Descriptive Statistics of Basic Variables
* Presence of high value-added statistical units
* Dublicate records

If there is a problem with the data then fully or partially re-transfer of troubleshooted data may require. If there is no problem, administrative data is matched and linked with the statistical business registers and reference tables including connection variables are created in order to enable subject matter units of TurkStat to be able to use it. Matching 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 and anonymization which is necessary for in-house use of individual data is performed at this stage. This is accomplished by anonymization of personal data 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.

*5.2. Data Processing*

Following the completion of the data acquisition phase, the process for processing these data is as follows:

* data integration,
* editing and checking for internal consistency,
* imputation and
* Finalize data for analyze phase ,respectively.

As regards the processing phase, data Integration is performed first by combining data from different tables, creating record linkage routines, deriving new variables and units (if necessary) and prioritization (in the case of data from two or more sources for the same variable). And then logical controls as well as internal consistency checks are carried out 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.

After all these steps, generated corporate data is then given access to TurkStat subject matter units for furher analysis.

**6. What Has and Has Not Changed In Business Statistics with the New System?**

With the integration of RA and SSI data sources, 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.

*6.1. Business Registers*

Thanks to the micro level use of RA and SSI records, important improvements in the BR have been realized. With the use of administrative data in business statistics, the BR started to be created on a daily basis. Accordingly, anytime within a year the business registers framework can be obtained. Additionally daily transactions can be added to the annual business registers framework.

*6.2. Short-Term Business Statistics (STS)*

Within the scope of STS; production, employment, hours worked, gross wages and salaries, labor costs, retail sales, producer prices indices and building permit statistics have been calculated for the industrial, construction, services and trade sectors for monthly and quarterly periods. In the past; only the building permit statistics generated from administrative data, and surveys were conducted for all other statistics. But now; along with monthly scheduled data acquisition from RA and SSI, short-term indicators have begun to be produced from administrative data with an exception of producer prices indices. Thus, forecast errors resulting from the use of fixed base year and limited number of observations are eliminated.

**Table 1. Short Term Business Indicators in the New Case**

|  |  |  |  |
| --- | --- | --- | --- |
| **New Series** | **Data Source** | **Period** | **Timeliness** |
| Turnover Indices | VAT and Special Consumption Tax Returns | Monthly | T+47 |
| Retail Trade Indices | VAT and Special Consumption Tax Returns | Monthly | T+49 |
| Industrial Production Index | VAT and Special Consumption Tax Returns, Monthly Industrial Production Survey | Monthly | T+47 |
| Labor Input Indices | SSI Monthly Premium and Service Return, Withholding Tax Return | Quarterly | T+60 |

*6.3. Annual Business Statistics*

Annual Industry and Service Statistics (AISS) aim at producing information that contributes to define enterprises according to their structural characteristics. The statistics produced by TurkStat within the Structural Business Statistics constitute the main components of the product structure and the economic structure that feed the system of national accounts in particular. AISS will be produced based on administrative records from this year. As a result of this; the application of AISS 2016 survey, which had been planned to be carried out with about 180,000 enterprises and imposes a high level of response burden on the enterprises, has been terminated.

*6.4. National Accounts*

RA and SSI data at macro level have been started to be used in 2009 based GDP series that were published on 12 December 2016. With the production of annual and short-term business statistics data from administrative records, the STS indices and AISS data are used instead of macro level RA and SSI data in GDP estimates. Thus, in addition to increased coherence between annual, short-term, national accounts indicators, the response burden on enterprises will be reduced, resources will be saved and data quality will be increased.

**7. Conclusion**

In conclusion; use of administrative data in statistical production is expected to reduce cost and reporting burden of businessess and improve data quality. Besides, shifting data collection from primary to secondary will contribute to the development of the institutional culture in a positive way. It will improve internal cooperation and coordination between units as well as adding new skills and mindset to the staff. It will also allow Turkstat to establish strategic partnerships and alliances with administrative records holders.

Increasing the use of administrative records in the official statistics production process has led to the adoption of new approaches by TurkStat in the process of compiling, processing and publishing data. These approaches, which require close cooperation between administrative data providers and TurkStat, will improve the analytical capabilities of institutions concerned over time while improving the quality of official statistics and thus contribute positively to the quality of administrative records.

TurkStat adopts a data collection strategy that primarily uses secondary sources having sufficient quality. In the absence of them, primary data collection, i.e. directly from the respondents, is implemented as a last choice. The strengths and weaknesses of data derived from administrative data compared to survey-based data in terms of quality determine how these sources can be used more effectively. Administrative data directly or indirectly replaces field surveys or survey responses or both and they are currently used for creating or supplementing frames and data evaluation. These uses now penetrate most of the statistical domains and use of administrative data is expected to expand further in the future in TurkStat.

**8. 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].

Statistics Law of Turkey, <http://www.turkstat.gov.tr/UstMenu/yonetmelikler/StatisticsLawOfTurkey.pdf> ,[Visited: 14 May 2018].