Proceedings of Statistics Canada Symposium 2018
Combine to Conquer: Innovations in the Use of Multiple Sources of Data

Transactional Data Processing System

Agnes Waye, Serge Godbout, and Nathalie Hamel[1]

Abstract

Transactional data is becoming more commonly used as administrative data or in surveys. The richness and volume of the data allows the user to gain valuable insight and to conduct a more thorough analysis of trends. However, such large datasets with complex structures pose unique challenges in terms of data processing and estimation, and classic data processing methods require adapted solutions. At Statistics Canada, there is a gap in the statistical infrastructure to process transactional data. We have identified the need to develop a more robust system to process transactional data since a high level of flexibility is required. A transactional data processing system has been developed for transportation surveys, which include many surveys with transactional data. One survey has been integrated into this system so far (Fare Basis Survey) and gradually, other surveys from aviation, rail and trucking statistics programs will be integrated as well. This system implements steps from the process phase as identified in the Generic Statistical Business Process Model (GSBPM), including features such as data import, edit and imputation, data integration, balancing data, and estimation. This paper will discuss the definition and the specific characteristics of transactional data, how they are processed, lessons learned, challenges we faced, as well as future issues to resolve in the transactional data system.

Key Words: Data processing; Processing; Edit and imputation; Estimation; Transactional data; Data integration.

1. Introduction

1.1 Background

Transactional data is becoming more commonly used as administrative data or in surveys. The richness and volume of the data allows the user to gain valuable insight and to conduct a more thorough analysis of trends. However, such large datasets with complex structures pose unique challenges in terms of data processing and estimation, and classic data processing methods require adapted solutions. At Statistics Canada, there is a gap in the statistical infrastructure to process transactional data. In particular, the Transportation Statistics Program at Statistics Canada uses transactional data for official statistics and analytical products. Three examples of transportation surveys that use transactional data are the Fare Basis Survey, Aircraft Movement Statistics and Trucking Commodity and Origin Destination Survey. The Fare Basis Survey is a regular and comprehensive source of fare type-specific data on passengers, revenues, and average air fares (Statistics Canada, 2018a). Aircraft Movement Statistics provides estimates of aircraft movements in Canada; the data are used by Transport Canada and NAV CANADA for measuring the workload of air traffic controllers, aircraft activity on air routes and runway utilization (Statistics Canada, 2018b). The objective of the Trucking Commodity Origin and Destination Survey is to measure the commodity movements and the outputs of the Canadian trucking industry (Statistics Canada, 2017). Due to the lack of existing tools to process the transactional data for transportation surveys, a new processing system for transactional data has been developed. One survey has been integrated into this system so far (Fare Basis Survey) and gradually, other surveys from aviation, rail and trucking statistics programs will be integrated as well. This system implements steps from the process phase as identified in the Generic Statistical Business Process Model (GSBPM, see UNECE Statswiki, 2018), including features such as data import, edit and imputation, data integration, balancing data, and estimation. This paper will discuss the definition

[1] Agnes Waye, Statistics Canada, 100 Tunney's Pasture Driveway, Ottawa, Canada, K1A 0T6 (agnes.waye@canada.ca); Serge Godbout, Statistics Canada, 100 Tunney's Pasture Driveway, Ottawa, Canada, K1A 0T6 (serge.godbout@canada.ca); Nathalie Hamel, Statistics Canada, 100 Tunney's Pasture Driveway, Ottawa, Canada, K1A 0T6 (nathalie.hamel@canada.ca)

and the specific characteristics of transactional data, how they are processed, lessons learned, challenges we faced, as well as future issues to resolve in the transactional data system.

2. Transactional Data

2.1 Characteristics of Transactional Data

Transactional data can be considered to belong to the family of big data. It shares the 4 v's with big data: volume, velocity, veracity, and variety. Transactional data tends to come in extremely large datasets due to the high frequency at which the data is reported and the high number of transactions. The velocity of the data is also a special feature since the data is often transmitted very quickly. This is unlike traditional data (such as survey data) where sometimes there is a long lag between the creation of the data and the collection of the data. The veracity of the data refers to how the quality of the data can vary, either depending on the provider or over time. The variety describes the diverse range of formats that transactional data can come in.

Transactional data is usually collected at pre-defined frequencies (for example, daily or weekly). Examples of transactional data include financial data, like invoices, or logistics data, like travel records. Variables can be related to time (e.g. date), classification (e.g. types of commodities), or numeric information (e.g. revenue). The variables of interest for transactional data are often summed up to create targeted statistical information. For example, total sales in a month can be calculated by summing up the sales of all transactions for a month.

2.2 Comparisons between Transactional Data and Traditional Data

Here is an example of what transactional data looks like, using the Fare Basis survey as an example. Table 2.2-1 shows the total number of passengers for each air carrier. In the Fare Basis Survey, each air carrier provides us with a file every quarter with air coupons, which would look like Table 2.2-2.

Table 2.2-1
Carrier Records

Carrier	Total Passengers
A	100
B	200

Table 2.2-2
Transaction Records

Carrier	CityPair	Passengers
A	AB	40
A	AC	60
B	AD	100
B	AE	100

Transactional data can be more flexible than traditional data, meaning data that are collected from traditional surveys. Table 2.2-3 shows what traditional data would look like if we wanted to add passengers for each sector (international, domestic and transborder). As seen in the table, we added three separate variables for these passenger counts. In Table 2.2-4, we can see that the transactional data shows the same information as the traditional data without needing to add three new variables. Only one new variable, "sector" was added to the transactional records. This example illustrates how transactional data requires less new variables to show domain level information compared to traditional data.

Table 2.2-3
Traditional data

Carrier	Total Passengers	Passengers – International	Passengers – Domestic	Passengers – Transborder
A	400	100	200	100
B	400	300	40	60

Table 2.2-4
Transaction Records

Carrier	CityPair	Passengers	Sector
A	AB	100	International
A	AC	200	Domestic
A	AZ	100	Transborder
B	AD	300	International
B	AE	40	Domestic
B	AF	60	Transborder

Now, let us consider a scenario where we want to add a new sector, "foreign". In the traditional data (Table 2.2-5), we need to add a new variable that shows foreign passenger counts. However, in transactional data (Table 2.2-6), we don't need to add any new variables. It is sufficient to add records directly to the transactional data with a value of "foreign" for the sector variable.

Table 2.2-5
Traditional data

Carrier	Total Passengers	Passengers – International	Passengers – Domestic	Passengers – Transborder	**Passengers – Foreign**
A	400	100	200	100	**N/A**
B	450	300	40	60	**50**

Table 2.2-6
Transaction Records

Carrier	CityPair	Passengers	Sector
A	AB	100	International
A	AC	200	Domestic
A	AZ	100	Transborder
B	AD	300	International
B	AE	40	Domestic
B	AF	60	Transborder
B	DF	50	**Foreign**

In Table 2.2-7, consider the case where there is also a revenue variable for the transactions, as well as a new domain variable called isWeekend (with values of either 1 if the reported day falls on a weekend, 0 otherwise). If Table 2.2-7 is to be transformed into a table in the traditional format like Table 2.2-5, many new columns would need to be added for each combination of isWeekend and Sector for passengers and revenue.

Table 2.2-7
Transaction Records

Carrier	CityPair	Passengers	Sector	Revenue	isWeekend
A	AB	100	International	100,000	1
A	AC	200	Domestic	300,000	0
A	AZ	100	Transborder	80,000	0

3. Details of the System

3.1 Framework of System

The functionalities of the processing system correspond to the "process" phase of the GSBPM. These steps include: data integration, validation, edit and imputation, variable derivation and calculation of estimates. The system consists of a set of tools to be integrated into new survey applications. The system is entirely SAS driven, and is made up of a set of SAS macros. There is a processor that runs all the steps. The system reads in parameters and steps from metadata, which can be customized by the user. If any parameters need to be changed, then only the spreadsheet containing the metadata needs to be updated. This system is modular and flexible and can be adapted to different survey processing models.

The system maximizes the use of corporate tools at Statistics Canada, such as BANFF for edit and imputation and G-Est for estimation (for more details on these 2 corporate tools, please see Statistics Canada, 2018c and 2018d). The system also includes a user guide.

Due to the modular nature of the system, all the steps can be reordered and repeated in any way, as long as the input and output files are correctly connected between steps. The tables below show the metadata spreadsheets that describe the processing steps. Table 3.1-1 includes the steps which should be executed and the order they should be executed in. Table 3.1-2 shows the parameter names and parameter values for each of the steps listed in Table 3.1-1. As seen in the tables, the steps can be reordered in any way and the parameters can be easily modified through spreadsheets. Each processing step produces its own set of logs and output, which facilitates debugging.

Table 3.1-1
Processing Steps

StepID	Module
1	Import
2	Import
3	stackFiles

Table 3.1-2
Processing Parameters

StepID	ParamName	ParamValue
1	inFileName	File1
1	outFileName	File1_out
2	inFileName	File2
2	outFileName	File2_out
3	listFilenames	File1_out file2_out
3	stackedFileName	File1File2Stacked

The two tables above (Tables 3.1-1 and 3.1-2) show a basic example of how steps can be executed using a processing system composed of two macros, one to import and format a file (called Import, with two parameters) and one to stack multiple files (called stackFiles, with two parameters). The processing steps table (Table 3.1-1) specifies the sequence of steps to be completed, which include the import step (twice) and then stacking files. The processing parameters table (Table 3.1-2) shows the details of the steps. First, a file named File1 is to be imported, with the output name being File1_out. Then, a file named File2 is to be imported, with the output name being File2_out. Finally, File1_out and File2_out will be stacked together to create a new file called File1File2Stacked.

As mentioned earlier, the system was developed to process transportation transactional data. Only one survey has been processed using the system so far, the Fare Basis Survey. This is a survey that collects revenue and passenger counts from air carriers to measure average air fare.

4. Challenges

4.1 Volume

The large size of datasets caused processing time to be extremely slow. Any errors took a long time to detect since each run of the program took a significant amount of time. Our recommendation is to minimize the file sizes by dropping unnecessary variables and to maximize the efficiency of the system by minimizing the number of steps. The current system includes steps right before edit and imputation and estimation to drop unnecessary variables. It is important to do thorough initial testing with very large files to see what the limits of the system are.

4.2 Integration of Data

When working with transactional data, often we have to combine datasets from different sources with different levels of detail, frequency and quality. It is important to first find a common layout, format and definitions before processing the data. The lesson that was learned is that it is important to build the processing model towards the output estimation files, and not the input file. One should start with the output tables, and then work backwards to figure out processing strategies of the input files.

4.3 Design

Since the system prioritized the requirements of Fare Basis Survey, it has been developed toward a 2 stage sampling design at the moment. In the Fare Basis Survey, stage 1 is a census of carriers, and stage 2 is a sample of transactions from carriers for selected days. We had to ensure there was coherence between the stages and to accommodate different designs for each stage. In the future, we have to work on developing more features for other designs.

There were also issues related to coverage. First, there was the issue of duplicate transactions. It is possible that different companies can report the same transactions or there can be transactions that cancel each other out. There is also the issue of missing transactions. It is sometimes difficult or impossible to know if transactions are missing. Transactions can be compared with historical data but it is still hard to determine with certainty which ones are missing.

The lesson learned here is that it is important to use benchmarking methods to adjust transactional data to external control totals (e.g. from other surveys or administrative files). It is important to work closely with subject matter experts and data providers to ensure a higher quality of data. Finally, it is important to correctly define what transactions are in scope from providers.

4.4 Estimation

As mentioned in Section 4.2, the system prioritized the design of the Fare Basis Survey. If there is a census at the first stage, then the system can calculate annual estimates. One limitation of the system is that it can handle annual estimation only if the same design is implanted for every sub-annual reference period (e.g. every quarter). We will need to develop extra features to accommodate different sub-annual reference period designs in the future.

Another estimation challenge is related to the issue of inactive units. For example, in the Fare Basis Survey, we may get a carrier that is actually out of scope for a specific quarter. Since transactions are used for estimation, all inactive units must be represented on the transactions dataset in order for variance estimation to be calculated correctly.

4.5 Imputation

There are also special considerations related to imputation when working with transactional data. Domain variables need to be given special attention when using historical imputation. If the domain variable is related to variables of interest, it is important to keep the historical value of the domain variable. For example, in the Fare Basis Survey, there is a variable called isWeekend defined in Section 2.2. When using historical imputation, we add one to the year of the reporting date, but we keep the historical value of the isWeekend variable. The reason for this is that the revenue

variable, which is a variable of interest, is strongly related to isWeekend (weekend fares tend to be more expensive). Therefore, it is important to preserve the historical value of the domain variable.

In the example below, Tables 4.5-1 and 4.5-2 show a partial calendar for January of 2017 and 2018, respectively. January 1 was on a Sunday in 2017, so its value for isWeekend is 1. However, if 2017 data is used to impute for 2018 data, we would keep the historical value of isWeekend for January 1 even though in 2018 it falls on a weekday.

Table 4.5-1
January 2017 calendar

January 2017						
Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
1	2	3	4	5	6	7

Table 4.5-2
January 2018 calendar

January 2018						
Sunday	**Monday**	Tuesday	Wednesday	Thursday	Friday	Saturday
	1	2	3	4	5	6

Unique identifiers also have to be taken into account when performing historical imputation for transactional data. In the BANFF edit and imputation system, it is a requirement for current and historical datasets to have matching unique identifiers in order to perform historical imputation correctly. However, for transactional data, we usually don't have matching identifiers for the two datasets. One solution is to match by domain groups instead of unique identifiers when performing historical imputation.

5. Conclusion

This paper described the components of our transactional data processing system, as well as the challenges and lessons learned when working with transactional data. There still remains a lot of work to be done to further refine, improve and expand the functionalities of the system. There will be different functionalities to be developed depending on the requirements of the survey being integrated. There will be 2 more surveys in the near future that will be processed using our system: Aircraft Movements Statistics and Trucking Commodity Origin and Destination Survey.

References

Statistics Canada (2017), *Trucking Commodity Origin and Destination Survey*.

Statistics Canada (2018a), *Fare Basis Survey*.

Statistics Canada (2018b), *Aircraft Movement Statistics*.

Statistics Canada (2018c), *BANFF (Edit and Imputation- Generalized System)*.

Statistics Canada (2018d), *G-Est (Estimation – Generalized System)*.

UNECE Statswiki (2018), *GSBPM v5.0*.