Chapter 5

Destination matrix building and disaggregated choice modeling using tax revenue data

Rodrigo Javier Tapia
Universidade Federal do Rio Grande do Sul, Porto Alegre, Rio Grande do Sul, Brazil

Highlights

- Lack of data, limed knowledge, experience, and funding for data collection and data handling.
- Development of hybrid approaches that combine from different public authorities.
- Successful application of the approach for the development of origin–destination matrixes and mode choice models.

1. Introduction

Data availability in the freight sector has been pointed out as one of the major causes for the slower development of the modeling techniques, compared with passenger transport (Tavasszy and de Jong, 2013). The quality and availability of data are important factors that affect not only the level of detail of the model but also the type of models that can be used (Chow et al., 2010). For example, if only aggregated data on market shares without any information on individual choices are available, a modal choice model will have a limitation on its predictive and interpretative value.

Data in freight are generated by the commercial relationships between firms. The records generated by these transactions hold sensitive information to their core functioning. This results in the reluctance to share data because of the risk of losing competitive advantage.

Data asymmetry could be responsible for the difference in the models of developed and developing countries. In the former, there are more

Freight Transport Modeling in Emerging Countries. DOI: https://doi.org/10.1016/B978-0-12-821268-4.00005-8

consistent, compatible, and comparable data sources, such as EUROSTAT, and more specific projects and data collection, such as the Commodity Flow Survey for the United States. Nevertheless, data sources are far from perfect, as they are not always periodically updated and are not always intended for transport modeling. For example, the French flow survey (ECHO) of 2004–2005 is still used for new models and applications (e.g., Jensen et al., 2019).

In emerging countries, although it is difficult to generalize, data are harder to collect. Sources tend to be less reliable, with shorter and not always consistent time series. In some cases, contradictory data can be found even if they all come from governmental sources. This is aggravated by the lack of clarity in the method used for collection.

Another issue found at this level is the lack of compatibility among different databases. As statistical records of different organizations were created and developed independently, they have different nomenclature and levels of detail that make data compatibility a much more difficult job.

Although the unstructured data are a big problem, it can bring some opportunities. Not having structured methodologies to construct and update data can give the chance to innovate in data collection and processing by generating new and more suitable models for the needs of developing countries.

These opportunities have not only appeared for freight transport. In passenger transport, big data and ubiquitous data have good prospects and several challenges. Ubiquitous data, generated independently and asynchronous from many different sources (Hotho et al., 2010), need further treatment to fill the gaps derived from the passiveness of its generation. This process, also known as data augmentation, is crucial to make the data usable for modeling.

Data augmentation is the process of obtaining new data to compliment an existing data source. This can be made by exploiting the current database or by relying on other sources to compliment it. For this stage, it is important to know the type of model to be used to gather the appropriate data. For instance, if the objective is to investigate traffic allocation, network and origin–destination (OD) data are needed; and if the objective is to build disaggregate behavioral models, data about the original choice situation have to be collected.

For example, detailed warehouse location data for developing a freight transport management assessment for the rice industry in Vietnam have been used (Binh, 2017). This level of detail is difficult to find, especially where data exhibition and collection are more standardized, the case for developed countries. Other data sources with high potential come from tax agencies. Databases from these sources are stable and structured because of their role to discover and control tax fraud schemes.

The need to maintain the tax secrecy poses the greatest challenge for their application in transport modeling. Normally, the privacy issue is the reason why these data are not available. Nevertheless, with the proper measures to protect privacy, it can be shared from the government for modeling and policymaking proposals. This is the case of the electronic invoice (EI) in the state of Rio Grande do Sul, Brazil, and the consignment bills (CBs) for grain transport in Argentina.

In the context of the Transport and Logistics State Plan (PELT, for the abbreviation in Portuguese) of Rio Grande do Sul, information about the movement of products measured in monetary value was available. This information was used to build OD matrices to identify the greatest infrastructure bottlenecks.

The CB used in Argentina consists of the origin and destination of all grain products and was made to have a tighter control on agricultural goods movements. It is a unique dataset to help understand the behavior of one of the most important productive sectors of the country.

The objective of this chapter is to describe and discuss the use of tax agency data for the cases of Brazil and Argentina. The data available, the processing, results, and augmentation will be addressed in both cases. Finally, some considerations on how to set up collaboration strategies with tax agencies to obtain the data. The chapter continues with section 5.2 discussing in depth the case of Brazilian EI use for OD matrices, whereas section 5.3 will discuss the use of CB for OD matrices and behavioral modeling. Finally, section 5.4 will show the conclusions of the chapter.

2. Description of the data sources

2.1 Electronic invoices

Brazil is the biggest country in Latin America in terms of population, gross domestic product (GDP), and size. Because of the federal nature of the country, several responsibilities fall on each state's government. One of these is transport infrastructure development and planning. In this section, the PELT of Rio Grande do Sul is addressed. In particular, the development of OD matrices from taxation data and its further validation is described.

Rio Grande do Sul is the southernmost state in Brazil. It borders with Argentina to the West, Uruguay to the South, and the Brazilian state of Santa Catarina to the North. The economic structure is diverse, ranging from agricultural production (for internal consumption and export) to the industrialization of heavy machinery.

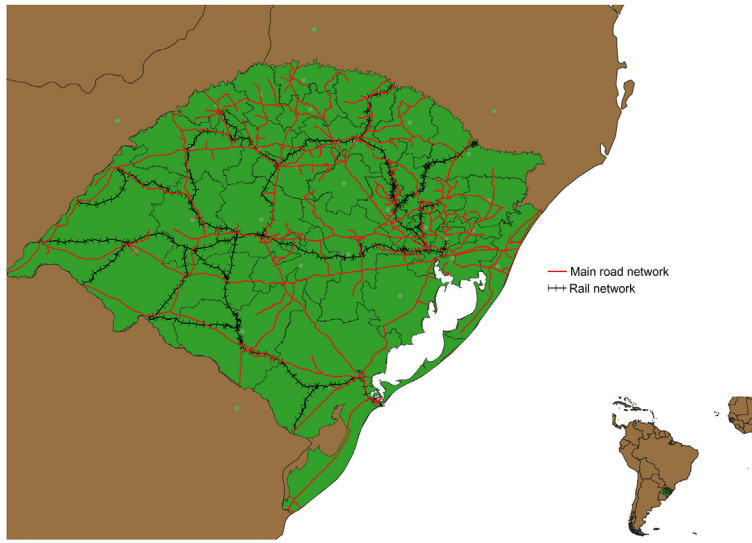

FIGURE 5.1 Main roads, railways, and waterways in Rio Grande do Sul.

The products are moved mainly by road, with 85% of the market share. Of the entire road network, only 8% is paved; almost all the roads under federal administration are paved; about 60% of the roads under state administration are paved; and about 98% of local (municipality level) roads are unpaved. The railway networks and the waterways are mostly oriented toward the export of unprocessed commodities to the port of Rio Grande, the only exporting port of the region. Fig. 5.1 shows the available transport network in the state.

The PELT had the objective of generating long-term infrastructure policy in concordance to national-level logistic plans. By establishing the baseline flows and forecasting them under various scenarios, infrastructure constrains were identified.

2.2 Description of the instrument

The tax collection instrument used by the PELT was the EI, created in 2005. Over the years that followed the introduction, the adoption of the EI for different segments of products and sectors was made mandatory. By the year 2010, most of the goods and services of the economy had to adopt the EI. By 2014, the year of the data collection of the PELT, its use was widespread, and the databases consolidated.

The EI had the objective of replacing paper invoices that documented the movements of goods or the provision of services. The result was

a simplified way of communication between companies and the taxation authority.

Data contained in the EI varied from product to product, but some information was common to all. These were data about the provider and client (location, name, and ID), the product (ID, quantity, and value), and taxes paid in the transaction.

2.3 Processing

To guarantee fiscal privacy, the data provided by the taxation agency of the state of Rio Grande do Sul [State Department of Taxation and Finance (SEFAZ), abbreviation in Portuguese] did not show the businesses involved, protecting taxpayers' confidentiality. Data were then aggregated at a macro zoning level. The 28 traffic zones adopted are consistent with the ones used by the Secretary of Planning, enhancing the interoperability of the information and results.

For origins or destinations outside Rio Grande do Sul, five extra zones were added. Two for Brazilian locations (Brazil East and Brazil West), one for eastern international flows that go through Argentina, one for southern international flows that go to Uruguay, and one for the flows that goes to the port of Rio Grande. In the latter, international export flows were confounded with coastal shipping to the rest of Brazil.

Another aggregation that was made by PELT had to do with the products involved. The EI has detailed information regarding the description of the product, something that is not necessarily needed in a long-term plan. Twenty-eight different product types plus one for general cargo were extracted. Only 5% of the movements, measured in Brazilian Reais, were discarded due to lack of consistency.

3. Consignment bills

Grain production is one of the most important economic activities in Argentina not only because of its participation in the GDP (10%), exports (70%), and tax revenue (10%) but because of the other economic activities related to the supply chain involved, such as crushing and agricultural machinery.

The grain supply chain can be divided into two, as shown in Fig. 5.2. On one side, there are the receivers, such as exporters, industry, or consumers. On the other side, there are the senders, such as producers and consolidators. Producers have two options for selling their crops. They can sell their products to exporters or industry (directly for the large producers, although brokers for the other producers) or to a consolidator. Because 46% of the production is carried by a large number of producers (Regunaga, 2009), they are likely to use the latter option. Besides, some large producers might also need conditioning for their seeds before selling

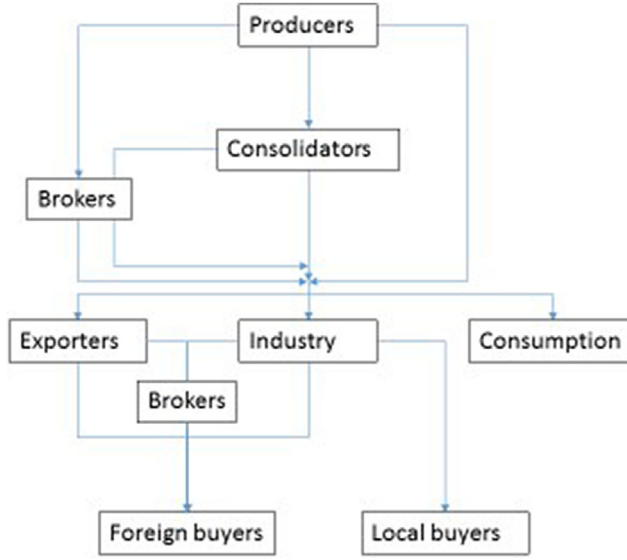

FIGURE 5.2 Grain supply chain. *Adapted from Instituto Nacional de Tecnología Agropecuaria (2009).*

them, and this is normally carried out by consolidators. Consequently, consolidators are a key agent in deciding where and how the inland transport is made in Argentina.

The mode choice of the grain is heavily unbalanced toward road. Around 84% of the volume is carried by truck, 14% by rail, and the rest by barge (Regunaga, 2009). Transportation represents an important share of the overall production costs. As a major factor in the competitiveness of the main exporting sector of the country, there is continuous interest in lowering logistic costs and increasing its efficiency.

Gaining a more in-depth knowledge of the transportation systems (and decisions) and the location of the production and demand centers is crucial to properly address this issue. Insights gained by OD matrices and behavioral models have an important role in this understanding.

3.1 Description of instrument

The CB for agricultural products originated as a way for the tax agency to control the commercialization of grain in Argentina. It works as proof of transfer of the grains. When it was first established, each of the participants of the transport (sender, receiver, and transporters) had a paper copy as a proof of the transaction. In 2009 the electronic support was implemented.

The CB had information on the sender (name and location), the transporter (name, mode, and tariff), the receiver (name and location), and the

TABLE 5.1 Data available from the CB.

Field	Data type	Description
ID	Integer	Line ID
Transport	Categorical	Mode of transport
ID of CB	Integer	CB ID number
Harvest year	Categorical	Year the grain was collected
Grain specie	Categorical	Type of grain transported
Origin ID	Integer	ID of the origin
Origin	String	Full name of origin
Destination ID	Integer	ID of destination
Destination	String	Full name of destination
Weight	Integer	Volume transported in kg
Unloading date	Date	Date of unloading

CB, Consignment bills.

product (type, weight, date of loading, and unloading). To protect the fiscal privacy of the organizations involved, some data were not shared. This consisted of information about the firms involved, prices, and date of unloading. Table 5.1 shows the structure of the shared data.

As the data recorded in the CB are about grain movement, each line represents an individual shipment. This means that if a sender has to break a bigger shipment into smaller ones to comply with the capacity constraints of a mode, this would mean that multiple records will appear in the database.

4. Origin destination matrix

OD matrices are that describe the movement of goods (or people in the case of passenger transport) between two zones. OD matrices allow transport modelers and planners to identify the main flows of goods between two pairs. The main differences between freight and passenger OD matrices are that freight flows do not always go through the shortest route between origin and destinations because of the need for the difference between inventory flows and production−consumption flows (Friedrich et al., 2014).

OD matrices are mostly used as an input for allocation models that can estimate the volume of flows through the networks. This use of the matrices is the objective of both applications in this chapter. The first one uses the EI to mount and calibrate the OD matrix for Rio Grande do Sul, Brazil, and the

second one shows the data augmentation process for building the matrices for grain transport in Argentina by using CB. This section is based on the reports of the works (Plano Estadual de Logistica e Transportes, 2015a,b; Ministerio de Transporte de la Nacion Argentina, 2017) and interviews with the coordinators of the projects (for the PELT that uses EI, the coordinator was Professor Luiz Afonso dos Santos Senna from the Universidade Federal do Rio Grande do Sul, and for the CB, the coordinator was Lic. Mariana Melgarejo, from the Ministry of Transport of Argentina), as the author of this chapter did not participate in their elaboration.

4.1 OD matrices for EI

The biggest challenge for using the EI for an OD matrix was to obtain the volume of the flows. Although in the data there was a column that referred to volume, it was not uniform across the entire database. Measurements such as units, boxes, or tons appeared in the data, among others. Even when tonnage was used, the reliability of the measure was considered to be relatively low because it was simply a declaration of quantity and was not necessarily the actual amount transported. To infer the quantities, some measure of the unitary value per ton had to be obtained.

Depending on the level of homogeneity of the group, different sources were available. For exported commodities that are typically homogeneous, a database for international trade was used (Ministério da Indústria, Comercio Exterior e Serviços, 2014). These records are a very reliable source of information of weights and monetary value because both amounts are recorded by the government and not declared by the exporter. This makes it a preferred source and used when available. The main issue was that the results obtained from international trade records might not represent the heterogeneity and relative participation of different products in the groups.

The second source used was the EI data itself. Because it relied on declared weight values, there was high variability on the amounts. Nevertheless, it contained enough disaggregation to absorb the heterogeneity present in each product category. A preliminary OD matrix with monetary values and another for weights was made to estimate the unitary value. After dividing both matrices, outliers were excluded before averaging them by product group.

Two commodity types were treated individually. The first one was vehicles and auto parts, where volume data from the EI were used directly. It was considered that because of the high value of the product, the measurements of volume found in the database were accurate enough. The second one was fertilizers, the main imported product in Rio Grande do Sul (in tons), where the average value and tonnage per truck were used.

Once the unitary value of all product types was obtained, the final OD matrices were estimated at a macro zone level. Traffic allocation at a macro

level has a lot of problems regarding the aggregation of origins and destinations because all flows concentrate on those points. This causes traffic to get assigned heavily into few routes, losing realism.

Additional data sources were used for further validation of the matrices and projected flows. OD surveys and traffic counting were used to disaggregate further the data from the EI. With 19,000 surveys distributed in 60 points of the motorways, factors that proportionally distribute intra- and extra-traffic zone flows were calibrated and applied (Plano Estadual de Logistica e Transportes, 2015a). Eqs. (5.1) and (5.2) show the estimation of the factors of attraction and production.

$$Fa_{iz} = \frac{v_{ai}}{\sum_z v_{az}} \tag{5.1}$$

$$Fp_{iz} = \frac{v_{pi}}{\sum_m v_{pm}} \tag{5.2}$$

where Fa_{iz} is the attraction of subzone i that belongs to the macro zone z, Fp_{iz} is the factor of production of subzone i belonging to macro zone m, and v_{ai}/v_{pi} are the volumes registered with destination/origin in subzone i. Eq. 5.3 shows the final decomposition of the volume between macro zone z toward macro zone m $\left(V0_m^z\right)$ into the volume from micro zone i to j $\left(V1_i^j\right)$.

$$V1_i^j = V0_m^z * Fa_j^* Fp_i \tag{5.3}$$

Each set of factors was estimated at a macro zone level, so the interpretation of them is how strong a district attract/generate cargo. An alternative to model this was to calibrate a gravity model to take into account transport impedances. The model proposed is a simpler one that only recognizes the relative importance of a district rather than the nature of the transport. With this method, the original OD matrices of 33×33 were decomposed into a matrix of around 500×500.

Additional traffic counting was made in 289 sections of the roadway for a week each. From these values, an approximate of the volume transported in those months was estimated. With the information of the toll posts, the seasonality effect was taken into account to annualize the volumes. Finally, the OD matrices were corrected by modifying the original OD matrix to fit the traffic counts. The method consists of minimizing the residuals between projected traffic and the traffic counts. Proportional adjustments are made to the seed OD matrix in an iterative way until a convergence rule is satisfied (Nielsen, 1998). In Fig. 5.3, the traffic allocation is shown for the production of soy.

This last validation procedure was made only for the baseline year. For the following years, the SEFAZ provided the same database of EI aggregated at a district level. This provided the information necessary for the allocation without having to rely on traffic counting and surveys.

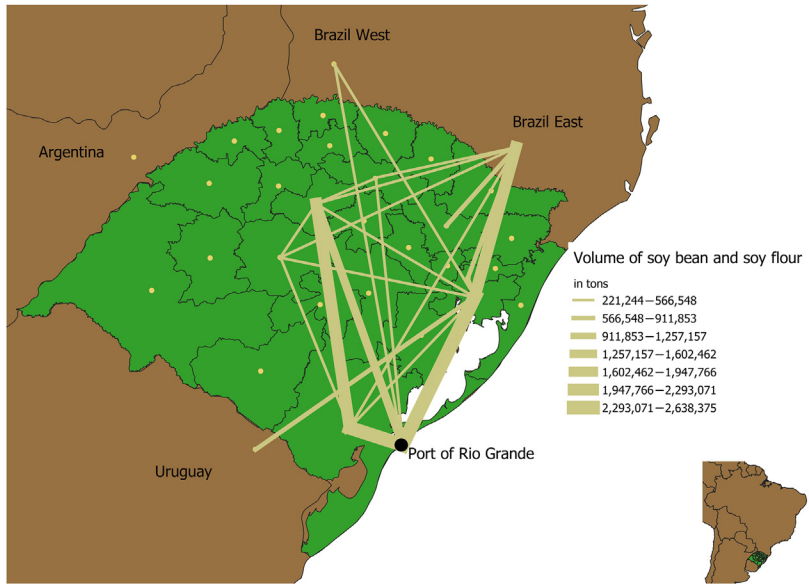

FIGURE 5.3 Origin and destination of soy.

OD matrices elaborations were the first step in building the models needed to identify infrastructure priorities. The expansion of the matrices was made following economic activity indicator's forecast, but the analysis of this model does not belong to the scope of the current chapter.

5. OD Matrices for CBs

The most straightforward application of the CB is to obtain OD matrices per product and mode. The results for all goods present in the CB and the main products in Argentina are published by the Ministry of Transport of Argentina (Ministerio de Transporte de la Nacion Argentina, 2017).

To obtain the product volumes, simple database processing such as grouping by OD pairs was made. For this particular application, the country was divided into 123 zones according to production profiles, population size, and other economic variables (Benassi, 2015). This zoning, although aggregated, gave the possibility to use other data sources and to compare results with previous works.

Some additional data cleaning was made to eliminate outliers. Several shipments with OD pairs without any commercial logic or very small shipments were excluded from the data.

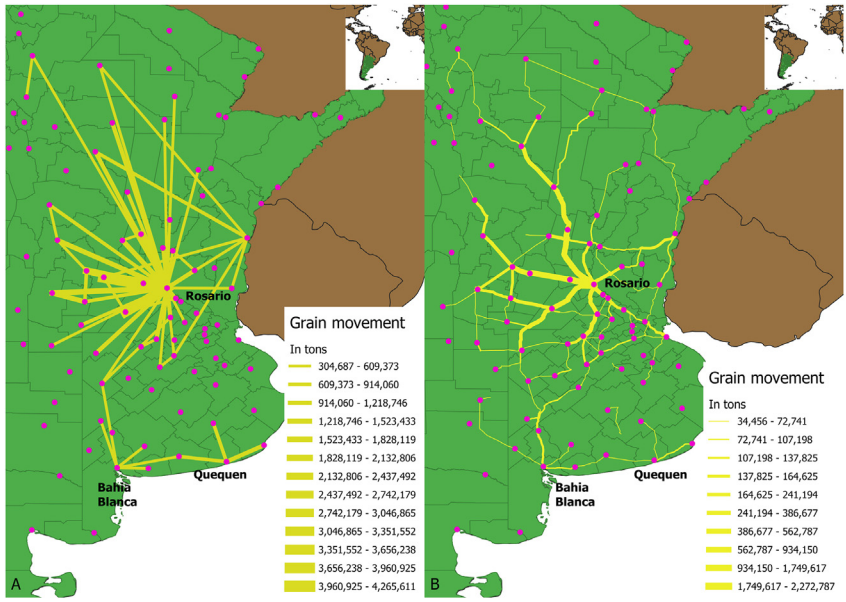

FIGURE 5.4 (A) OD of grains in Argentina. (B) Allocation of grain movements to the road network.

5.1 Data augmentation

For this application, few external data sources were needed. The main external inputs were related to zoning and aggregation aspects.

Regarding the allocation of flows to the existing network (one of the main objectives of the study), data on the network were used. It consisted of geolocated data of the main roads and railways that connected the centroids of the 123 zones. As only the main links between the centroids were used, the minimum path algorithm was used. Because of the lack of supporting traffic counts, only the volumes at the port were validated. The results for grain movement are shown in Fig. 5.4.

6. Mode choice

For many years, freight demand modeling relied on aggregate models, mainly because of data scarcity. Disaggregated models consider each decision as an individual choice. Modes yielded at disaggregated levels are more compliant with microeconomic theory.

Random utility maximization (RUM) is the theoretical background of most disaggregate behavioral models (Domencich and MacFadden, 1975). It assumes that decision makers try to maximize their utility in every choice. The utility for mode i for the choice maker q is given as follows:

$$U_{iq} = V_{iq} + \varepsilon \tag{5.4}$$

where V_{iq} is the observed utility, and ε is an error term. It is in the V_{iq} where the modeler characterizes each mode. The attributes could be cost, time, and reliability, among others. RUM establishes that the alternative with the highest utility is the chosen one.

Assumptions on the error term dictate the model form. If ε follows a type I extreme value distribution, the probability $P(i)$ of an alternative i being chosen among the choice set j is given by the multinomial logit:

$$P(i) = \frac{e^{V_i}}{\sum_j e^{V_j}} \tag{5.5}$$

It can be seen from the equation which data are needed for modeling choices. Four categories of data are required. First of all, data on the choice made are required. The second group is the choice set the individual is faced with. In other words, which alternatives are available to be chosen. The third category is the attributes that characterize each alternative (such as time, cost, and reliability). Finally, information on the choice maker (or shipment) can also be used to enrich the model.

Additional model structures are also available that can give a better understanding of the choice and provide better models. Some examples are the nested logit, cross nested logit, ordered logit, and probit.

Difference among choice makers can provide useful insights into their behavior. Heterogeneity can be introduced with random components in a mixed logit or with observed heterogeneity, such as socioeconomic variable interaction or latent class modeling.

6.1 Electronic invoices

Modal choice was not modeled in the PELT with EI data because of the aggregated nature of the data. Modal choice was addressed through an stated preference survey to identify the preferences of freight logistics managers to encourage intermodality and reduce overall logistics costs (Larranaga et al., 2017). It is considered to use the EI data in the future for this application.

6.2 Consignment bills

This database of individual shipments had the potential to be used for modeling disaggregate modal choice. To use it, the original choice situation had to be recreated. Owing to the characteristics of the privacy concerns the data had, some of the parameters of the choice situation had to be inferred. The process of data preparation and processing used in Southeast Buenos Aires Province for behavioral modeling (Tapia et al., 2019) is described in this subsection.

6.3 Data processing

The main challenge was derived from the structure of the data and its original objective. The CB was created to track individual grain movements, so each record can actually be part of a larger choice. In the hypothetical case of a shipment of 300 ton of soy, if it were going to be transported by truck, because of capacity constraints, it would have to be divided into 10 trucks. As a consequence, 10 different CB would be issued, and thus 10 different records would appear in the data. Nevertheless, only one decision was made. Not taking into account this issue may lead to an overrepresentation of road share. This issue is less likely to appear in other modes because they can transport larger loads.

To amend and remake the choice situation, multiple shipments had to be merged. It has been done by assuming that consecutive records from the same location, product, destination, and date come from the same larger shipment. The main underlying assumption is that when the sender fills the electronic form, it orders more than one at the same time, so consecutive numbers are issued.

The assumption is broken if the sender decides to add an extra truck *a posteriori* or if somebody else fills a CB with the same characteristics immediately after. The former would underestimate shipment sizes because they would exclude shipments from the same decision. The latter, although much more unlikely to occur, would overestimate the size of the shipments because of the over inclusion of records. After this processing, the database went from 2,932,686 records to 1,104,243.

6.4 Data augmentation

The compacted database has information on which mode was chosen, the (estimated) size, date, origin, and destination. This has no data on the choice itself or the not chosen alternatives, so additional information was gathered to recreate the full choice situation.

The first step was to decide the choice to be modeled. It could be simply mode choice, destination choice, or a mixture of both. In this case, the joint choice of mode and port was modeled. Consequently, the choice set was going to be all the combinations between ports and modes. Fig. 5.5 shows the main export ports, the rail network, and the productive area of Argentina. The products to be included in the analysis were the main exported agricultural products because of its importance in Argentina's economy. These were soy, wheat, sunflower, and maize.

There are three main ports in Argentina: Rosario, Bahia Blanca, and Quequén. Rosario consists of multiple terminals located alongside the Paraná River. The multiple terminals cover over 46 km of river banks and other cities, such as San Martín and San Lorenzo. It moves approximately 68% of all

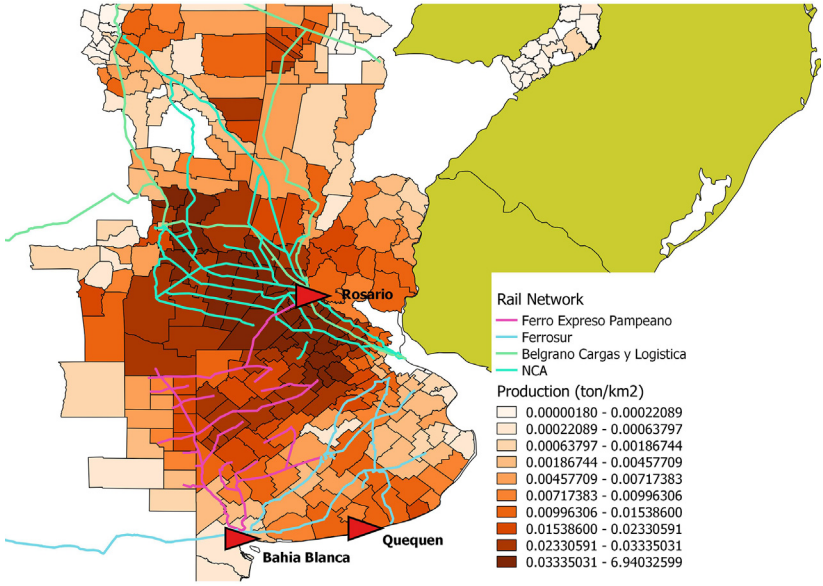

FIGURE 5.5 Main agricultural productive area and ports of Argentina.

grain exports of the country, and its price is normally used as a reference for grain commercialization. Most crushers and grain-related industries are located in the surrounding area.

By grain volume trade, Bahia Blanca is next in size. Located in south Buenos Aires Province, it is the deepest port of Argentina and goes directly to the Atlantic Ocean. For some trade routes, ships that are loaded in Rosario come afterward to Bahia Blanca to be completed. There is also a strong crushing industry, alongside with petrochemical manufacturing.

Another grain port worth mentioning is Quequén, in the southeast Buenos Aires Province. It is a port located in a river mouth, making it dependable on river ties for having an operative draught. Multiple operational issues have caused the prices paid in this port be systematically lower than other ports. Smaller ports, such as Ramallo (located on northern Buenos Aires Province, close to the Parana River mouth), were identified, yet discarded for analysis because of its low share of shipments.

The element chosen to characterize port attractiveness was the price paid for the products at the port. In this context, the Free Alongside Ship price was chosen because of being a published price at each port, which is used as a reference for transactions. Each port publishes at the end of the day the average price each grain was negotiated. When a round (day) had no price, there were no sales at the port, so the reference used is of the last available. This means that with the incomplete series of prices, it is possible to fill the

gaps. Prices of soy, wheat, sunflower, and maize for the year 2014 were included (FyO, 2018).

Transport network information was necessary to obtain basic information about the conditions of the transport. The objective was to process the network data to produce distances, times, and costs. Georeferenced maps for rail and road in the format of shapefile were provided by the Ministry of Transport. The road network consisted of a complete representation of the primary, secondary, and tertiary ways. The latter were discarded for the analysis because they were the small rural paths.

The rail network used had some problems with its integrity. After the privatization of the railway, several lines stopped functioning, which was ill informed to the regulatory agency. A big effort of the agency that controls rail infrastructure was made to obtain the operational status of more than 25,000 km of rail tracks. In the shapefile, data about condition of the track, maximum speed, and length were among the information available. Nevertheless, some problems were found when using data such as inconsistent speeds or some broken links. This resulted in using this information as a reference for average speeds for rail rather than using the actual reports of the network.

A python package called "networkx" was used to calculate the distances using the Dijkstra algorithm to find the shortest path. One path for road and two for rail (minimum distance and minimum time) were extracted, to capture network differences and travel times. All the combinations of origins and port destinations were considered.

Truck pricing is regulated in Argentina, so the prices per ton km are published (CATAC, 2014). Nevertheless, this is not perfect information because it has been reported that larger companies pay below this price and that some consolidators have their fleet.

Railroad prices are not regulated, and most companies do not publish them online. The only price table found was for the public-owned company and was used as a reference for the prices of all the network (Belgrano Cargas y Logísticas, 2014). Different pricing strategies appear during the year, and it has been mirrored in the augmented data.

Additional information on the region could be deduced from the data. From the distance matrix, for each origin, the closest port is noted. In regions that are within the influence, more than one port can be identified. From the Ministry of Agroindustry, some regional profiles of the consolidators (e.g., the number of companies or number and size of storage facilities) were used to further describe the firms in a region (Ministerio de Agroindustria, 2016). This information can be used to capture some heterogeneity in the choices made in the data.

Some data about the level of service the rail can deliver to a given area could be inferred from the CB. Knowing that the rail does not provide an everyday service, and considering that there is a constraint in rolling stock

during high season, it can be said that the smallest day lapse between two rail shipments of the same region is the best headway the rail can provide to that area. To do so, a filtered data from train had to be sorted by location and then by date. After that, the minimum difference between two dates would be the value of the headway.

Considering that the date provided is the unloading date, it could happen that the same train is unloaded in different days. In that case, the processing would throw a very small and unrealistic value for the headway. To solve this, a minimum headway had to be established. The value adopted was of 5 days for this application and came from interviews with users of the system.

Another element that can improve the results in the choice models is to refine the availabilities for each choice. Allowing the possibility of an alternative that was actually not available in the choice set could lead to bias in the parameters estimated.

It could be inferred that if there was no record of any train service to a certain location, rail was not available in the choice set of that site. Something similar could be done with ports, although every port could theoretically be reached by truck. Further modeling could be made to refine this, as made for passenger transport (Calastri et al., 2017), but it escapes from the scope of the chapter.

Most of the data gathered to augment the CB data had information on characteristics of either the origin or the destination of the choice. The final step was to join the tables into a single one to recreate the choice task. Table 5.2 shows the data structure of the consolidated data.

By processing the CB and with the use of other sources, revealed preference (RP) data were generated that could be used to generate disaggregate modal choice models. Although it is very rich information, several assumptions had been made during data augmentation and processing. Therefore the definition of some variables could become less objective, bringing potential bias and problems in the estimation and interpretation of results.

6.5 Application

An example of the application is the modeling of the port and mode choice for grain consolidators. The case study is from a region between the two most important grain maritime ports: Bahia Blanca and Quequén, as seen in Fig. 5.5 (Rosario, is a fluvial port). This region is interesting for understanding the competition between both ports. The CB will provide RP data, meaning that the model would be suitable to estimate elasticities and to forecast future scenarios. The choice modeled was the mode (train or truck) and port choice (Bahia Blanca or Quequén).

To reduce the bias that could be introduced in the inference of the data augmentation, an SP experiment was done. SP has the advantage of being flexible and designed by the researcher, but it bases on hypothetical choices

TABLE 5.2 Augmented data for disaggregated choice modeling.

Field	Data category	Description	Source
FAS price	Attribute	One field per destination; price paid at destination	Stock market
Distance rail	Attribute	One field per destination; distance traveled measured in the rail network	Network
Distance truck	Attribute	One field per destination; distance traveled measured in the road network	Network
Freight cost rail	Attribute	One field per destination; time estimated through distance and cost per km	Rail company + network
Freight costtruck	Attribute	One field per destination; time estimated through distance and cost per km	Truck association + network
Headway	Attribute	One field per destination; minimum day separation between two shipments at the origin	CB
Time rail	Attribute	One field per destination; time estimated through distance and average speed	Network
Time truck	Attribute	One field per destination; Time estimated through distance and average speed	Network
Availability	Availability	One field per destination; dummy if destination is available for rail	CB
Choice	Choice	Mixed mode and choice used	CB
Grain species	Choice	Type of grain transported	CB
Shipment size	Choice	Volume of the shipment in kg	CB
Closest port	Choice maker	Hinterland the origin belongs to	Network
Installed storage capacity	Choice maker	Volume the origin can store in permanent facilities	Ministry of agriculture
Mixed hinterland	Choice maker	Dummy indicating if there are multiple port competing	Network
Date	Miscellaneous	Date of the unloading	CB
Dollar	Miscellaneous	Conversion rate ARS/US$	Stock market

CB, Consignment bill; *FAS*, Free Alongside Ship.

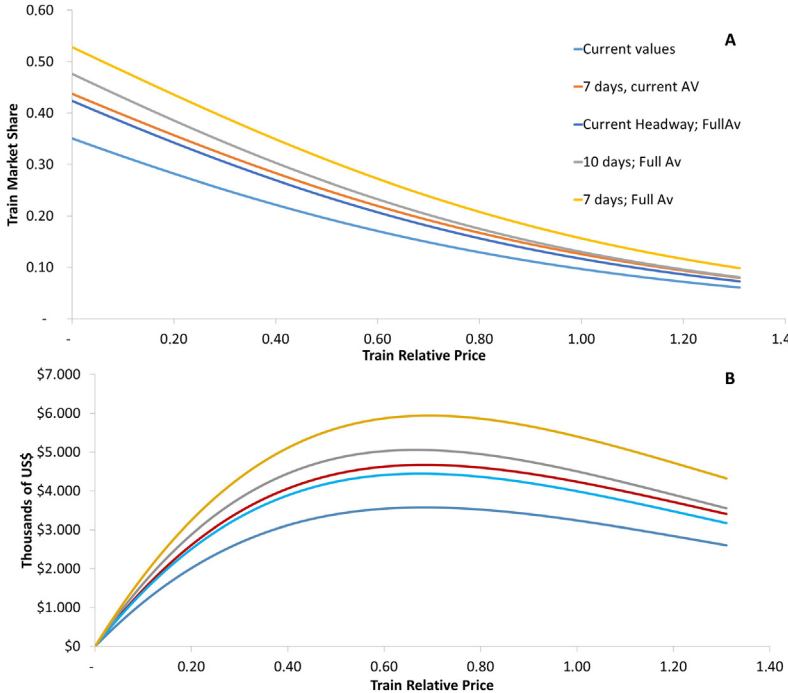

FIGURE 5.6 (A) Simulated train market shares; (B) Simulated rail incomes.

that might not be reflected in real-life scenarios. The biggest strength of the SP experiments in this context is that the variables are correctly defined, whereas with the CB processing, they are estimations of the actual variables. This way the SP can potentially correct some problems derived with the high correlation between the explanatory variables. By modeling RP and SP data simultaneously, the weaknesses of each method are overcome (Hensher et al., 2008). This results in a model that could be used for forecasting and elasticity estimation while having better properties in the variable definition. The details of the SP and the model results can be found in the paper.

The main outputs of the model were the simulation of the effects of improvements in rail services, different relative prices between train and truck (Fig. 5.6A), and optimal pricing strategies for rail (Fig. 5.6B). In addition, the port hinterlands for Bahia Blanca and Quequén and elasticities were estimated. The full estimations and results can be seen in the study by Tapia et al. (2019).

7. Discussion

The systematic collection of tax records brings a structured and consistent database about the "footprints" of the economic transactions. These

footprints correspond to the interaction between individual companies, and they dictate the economic relationship between the organizations that will derive in a freight flow.

The disaggregated (i.e., at an individual level) nature of the data brings opportunities and challenges. The opportunities revolve around obtaining good quality data for aggregated (e.g., OD matrices by aggregating the individual records) and specially disaggregated models (such as disaggregated mode choice). However, there are several challenges to be faced to make these data adequate for modeling.

The main challenge comes from the nature of the data: taxation. As they are fiscal data, before the data are available for modeling, they have to be anonymized to secure fiscal secrecy. The way that this is done has a direct influence on the models that can be used.

Within the restriction of fiscal secrecy, there are several levels of details that the data can have. The first level is obtaining aggregated data. The aggregation can be at a regional level (district, cities, provinces, etc.) and/or at a product level (can range from category of products or bundled in commodity types). This level of aggregation does not allow the modeler to use disaggregate models, but it does provide quality and trustworthy information of OD matrices and can potentially be used for aggregated distribution and allocation models. If there are previous demand models that use RP data, the new fiscal data can be used for recalibrating the models for the current modal shares. It has to be noted that flows that are unique for a company (i.e., only company of the commodity type in an area) can be omitted from the aggregation because it would violate the fiscal secrecy. This concern is shared by all levels of details.

The EI used in the PELT-Rio Grande do Sul presented in this chapter belongs to the first level of aggregation. The main application for the data was for building reliable OD matrices, but the application of behavioral choice models was limited and needed additional information to calibrate the models. The data were given at a higher level of aggregation than the open needed for the application. This meant that complementary studies needed to be done to obtain the level of detail required for the objectives of the PELT.

The second level of detail that can be obtained is removing the involved organizations in the transactions but showing the individual records. This allows to model actual flows in a disaggregated manner, for example, for a behavioral modal choice model. It is also easy to move from this type of individual data to the aggregated models mentioned before. Although it is a very interesting source of usually elusive disaggregate data, the lack of knowledge of the companies brings some limitations of the models: there is no knowledge of successive choices of the same company. This makes that any chained sequence of choices reflected in different documents are lost, together with possible changes in preference with time.

The CB belongs to this type of information. The data involved consist of the records of agricultural goods with the receiver, carrier, and shipper cropped. To solve the problem of chained choices, relevant to the choice situation here, the merging of the records commented in section 5.3 was made. It is worth noticing that this merging was able to be made because of the particularities of the tax instrument, and it is not really clear whether it is applicable in EI type of data.

The third level can involve more information on some of the companies involved by sampling and to give panel data from some of the companies and anonymized information for the rest. The panel data from companies must have an alias to maintain the secrecy and come from commodity groups and areas that make the individual identification impossible. Thus this additional level of detail is likely to appear in places with more density of producers of a certain good. Unfortunately, no study was found that uses these types of data.

Trust and cooperation between the taxing agencies with the modelers are crucial to determine the level of detail of the information. Models with tax data tend to appear after a collaboration with a governmental agency that can interpret and process the data itself or in collaboration with the academic or private sector. Moreover, the more the detailed the information is, the more responsibility and work does the tax agency has. In this subject, it is important that the modeling part is aware of the data structure and limitations and that it is specific with the data requests: the tax agency is the party with the highest responsibility and liability with a minimum reward for the results of the modeling. In all cases, data protection protocols are crucial to ensure that the data are used for the intended purpose and for the authorized parties.

To develop up this last item about the collaboration with tax agencies, the following recommendations are suggested:

- Establish the strategic importance of the models, have clear the objectives of the model, and be precise on how the tax information can improve the current models. This helps to get political support from the different areas of the government and also makes the tax agency to engage with the project.
- Be specific of the data format, information source, and preprocess needed. As the tax agency has to make sure the fiscal secrecy is maintained (it is their responsibility), they are not always available to do extra data work without being explicitly asked for.
- Be ready for a lengthy process. The time from the data request to when the data are available can be long and delay the main project if no data contingency plan is considered. Moreover, this time is mainly responsibility from the tax agencies, so there is little to be done externally.
- Expect to do plenty of data cleaning and processing. It is unlikely that the tax agencies provide any refinement of the information, so it is important to include time for data processing in the project's plan.

- Communicate the results. The application and model results can prove to be of importance to establish a more continuous flow of information for updating the model or for future efforts. If an effective and fluent collaboration is made, it is possible to obtain more detailed information in the future.

8. Conclusions

Within all the data problems that freight modeling has encountered, the use of nontraditional data sources is an important opportunity. This is especially true for developing countries because of the lack of structured data. In many cases, there are more nontraditional datasets available than in developed countries. One example is data that come from tax revenue agencies. The main concern with this source is that because of fiscal privacy, the access to the whole dataset is limited. Nevertheless, there is still valuable information that could be extracted with some processing and data augmentation. EI and CB are the cases shown in this chapter.

The EI has been used in the estate of Rio Grande do Sul, in Brazil, to help in the elaboration of the logistics master plan. The objective was to identify the main infrastructure constraints to prioritize infrastructure investment for the next 25 years.

The EI was used to create an OD matrix. The data, processed by SEFAZ to maintain privacy, consisted of monetary flows per group of products in macro regions. The processing consisted of transforming the monetary values into volume measures to standardize it and convert them into traffic for network assignment. Additional data collection was needed to decompose the aggregated matrix into district commodity flows.

CB was used in Argentina to control the grain movement. Two uses had been reported for freight modeling. The first one consisted of obtaining the OD of all grain flows. Additional inputs involved network data and information about macro zones.

The other use was to generate a dataset suitable for the estimation of disaggregated behavioral models. Several assumptions had to be made to make it compatible for choice modeling. The first one was to consolidate multiple shipments to recreate the choice situation. Further data augmentation consisted of obtaining data about the nonchosen alternatives, such as network analysis for distances, freight costs, and times. Finally, alternative assumptions for mode availability were made.

There are several levels of details that can be obtained from tax data. They limit and determine which models can be obtained from the data and the type of extra information needed. The greater the level of detail, the greater the effort and risk of the taxing agencies. To improve the likeliness of obtaining better data, it is important to have a proper definition of objectives and strengths of the model and to obtain a high engagement of the tax agency.

Although the results are promising, there are many limitations to overcome, mainly because of the assumptions caused by the privacy protection measures taken before the information sharing. This limits the reliability of the augmented data and can induce bias into the model estimation. Nevertheless, once the first procedures for processing are established, the specification of the data needed becomes clearer. The opportunity of being more specific in future data requests or even receive preprocessed data by the tax revenue agency appears.

References

Belgrano Cargas y Logísticas, 2014. Tarifario.

Benassi, A., 2015. Una Matriz Origen−destino Para El Transporte de Cargas En Argentina. PAMPA 12, 307−329.

Binh, N.T. 2017. A multi-stage impact assessment method for freight transport management measures Technische Universität Darmstadt.

Calastri, C., Hess, S., Choudhury, C., Daly, A., Gabrielli, L., 2017. Mode choice with latent availability and consideration: theory and a case study. Transportation Res. Part. B: Methodol, 0, 1−12. Available from: https://doi.org/10.1016/j.trb.2017.06.016.

CATAC, 2014. Confederación Argentina Del Transporte Automotor de Cargas: Tarifario. 2014.

Chow, Y.J., Yang, C.H., Regan, A.C., 2010. State-of-the art of freight forecast modeling: lessons learned and the road ahead. Transportation 37 (6), 1011−1030. Available from: https://doi.org/10.1007/s11116-010-9281-1.

Domencich, T., MacFadden, D., 1975. Urban travel demand. A behavioral analysis. North Holland Publishing Company, Amsterdam.

Friedrich, H., Tavasszy, L., Davydenko, I. 2014. Distribution structures, In *Modelling freight transport*.

FyO, 2018. Precios Históricos de Pizarra de Soja, Trigo y Maiz [WWW Document]. Available from: https://news.agrofy.com.ar/granos/precios/series-historicas/pizarra (accessed 25.10.18).

Hensher, D.A., Rose, J., Greene, W.H., 2008. Combining RP and SP data: biases in using the nested logit 'trick'-contrasts with flexible mixed logit incorporating panel and scale effects. J. Transp. Geogr. 16 (2), 126−133. Available from: https://doi.org/10.1016/j.jtrangeo.2007.07.001.

Hotho, A., Pedersen, R.U., Wurst, M., 2010. Ubiquitous data, *lecture notes in Computer Science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics)* 6202 LNAI: 61−74. https://doi.org/10.1007/978-3-642-16392-0_4.

Instituto Nacional de Tecnología Agropecuaria, 2009. Análisis de La Cadena de Soja En Argentina." Proyecto Específico 2742: Economía de Las Cadenas Agroalimentarias y Agroindustriales, 2009.

Jensen, A.F., Thorhauge, M., de Jong, G., Rich, J., Dekker, T., Johnson, D., et al., 2019. A disaggregate freight transport chain choice model for europe. Transportation Res. Part. E: Logist. Transportation Rev. 121, 43−62. Available from: https://doi.org/10.1016/j.tre.2018.10.004.

Larranaga, A.M., Arellana, J., Senna, L.A., 2017. Encouraging intermodality: a stated preference analysis of freight mode choice in Rio Grande Do Sul. Transportation Res. Part. A: Policy Pract. 102, 202−211. Available from: https://doi.org/10.1016/j.tra.2016.10.028.

Ministério da Indústria, Comercio Exterior e Serviços, 2014. AliceWeb, 2014.

Ministerio de Agroindustria, 2016. Infraestructure: Acopio y Almacenaje, 2016.

Ministerio de Transporte de la Nacion Argentina, 2017. Matrices Origen-Destino de Transporte de Carga 2014. https://datos.transporte.gob.ar/dataset/informe-matriz-origen-destino-vial-de-transporte-de-cargas.

Nielsen, O.A. 1998. Two new methods for estimating trip matrices from traffic counts travel behaviour research: updating the state of play. https://doi.org/10.1016/B978-008043360-8/50013-3.

Plano Estadual de Logistica e Transportes 2015a. Modelagem, http://www.pelt-rs.seinfra.rs.gov.br/index.php/andamento.

Plano Estadual de Logistica e Transportes 2015b. "Situação Atual: Conclusão." http://www.pelt-rs.seinfra.rs.gov.br/index.php/andamento.

Regunaga, M. 2009. The soybean chain in Argentina, January: 1–32.

Tapia, R.J., dos Santos Senna, A.L., Larranaga, A.M., Cybis, H.B., 2019. Joint mode and port choice for soy production in Buenos Aires Province, Argentina. Transportation Res. Part. E: Logist. Transportation Rev. 121, 100–118. Available from: https://doi.org/10.1016/j.tre.2018.04.010.

Tavasszy, L., de Jong, G., 2013. Modelling freight transport. Elsevier.