Australasian Transport Research Forum 2018 Proceedings

30 October – 1 November, Darwin, Australia

Publication website: http://www.atrf.info

# Understanding bid-price generation for road freight transport by analysing online market data

Hendrik Braun[1], Doina Olaru[1]

[1]University of Western Australia Business School, 35 Stirling Highway, Perth WA 6009

Email for correspondence: doina.olaru@uwa.edu.au

## Abstract

This paper aims to improve the understanding of road freight transport price generation by analysing quantitative and qualitative data already available. The online freight marketplace in Australia is used as a case study and a mixed-methodology adopted to obtain insights from qualitative information collected around bid-price generation, to identify patterns for freight transport and predict prices for palletised road transport.

The results suggest three clusters of consignments, each with slightly varying predictors of the bidding price. Although traditionally transport operators based their prices on distance alone, many other factors (customer type, flexibility, loading facilities) could be easily incorporated in the bidding model.

The paper shows how mixed-method research on publicly available data can contribute to more competitive decision making in daily operations. To the best of the authors' knowledge this type of research has not been undertaken before when analysing freight transport.

## 1. Introduction

In Australia, small operators, contributing 85% to the overall market revenue, dominate the road freight industry. Understanding the structure of the market is essential in order to make robust demand management decisions that are competitive and profitable. Yet, specific information for the industry, especially by region or type of commodities, is limited. Most of the published data is aggregated within government reports, or papers and fact sheets published by other organisations (e.g., Bureau of Infrastructure Transport and Regional Economics, BITRE, 2003 & 2014; Gargano, 2014a-c; KordaMentha, 2012). Information on the overall freight market structure (i.e., operator structure, types of goods, main transport routes, etc.) is also provided by the Australian Bureau of Statistics (ABS). However, detailed supporting material needs to be obtained from other sources, primary and secondary.

This paper reports on the use of secondary road freight transport negotiation data to better understand transport rate determination and its influencing factors in Australia, using web mining. The proposed mixed-method approach aims to reveal both numerical factors and shipper carrier interaction, which is often omitted, and may play an important role in understanding freight rates.

Research data is sourced from www.truckit.net, a meeting-place website for customers and freight operators where the former include their requests and the latter bid for jobs, based on their projected costs and capacity. If a customer accepts the bid, the request becomes integrated in the current routing of the 'winning' company, otherwise new bids may be submitted. Substantial data is available on the website, however it remains largely underused by industry, as companies are more concerned with pursuing potential customers, rather than understanding overall market characteristics.

In order to leverage this data source, parts of it (palletised transport) have been selected and extracted. The text segments were processed using a content analysis tool, Leximancer; then the numerical records, including several qualitative variables, were analysed using multivariate techniques. Three clusters/patterns of shipments emerged, and for each cluster, the dependence of the last bid on the characteristics of the demand and the transport conditions were explored. This last stage provides insights on the pricing strategy adopted by the freight companies.

The structure of the paper is as follows:

Section 2 presents scholarly work on web mining and methodological approaches, Section 3 details the implementation of the web content mining, which is followed by the data analysis (both qualitative and quantitative) in Section 4, showing the main determinations of the price formation. A discussion of the results, limitations, and ideas for further exploration concludes the paper (Section 5).

# 2. Literature review

Given the relatively reduced use of web mining in freight transport modelling, we deem appropriate a brief introduction of this technique, which has the potential to enhance data collection efforts, non-intrusively, at minimum cost.

## 2.1 Data mining

Data mining refers to a systematic analytical process to search, retrieve and convert various types of data into meaningful information (Witten *et al*., 2016). This is done by seeking patterns or regularities in substantial amounts of data and using that knowledge to produce more accurate future predictions. Han *et al*. (2011) use the term 'knowledge mining from data' which involves several steps (pp.7-8): 1) data cleaning (removing noise and inconsistent data); 2) data integration (combining multiple data sources); 3) data selection (retrieving data relevant to the analysis task); 4) data transformation (consolidation of the data into forms appropriate for mining); 5) data mining (applying intelligent methods to extract data patterns); 6) pattern evaluation (identifying those interesting patterns representing knowledge); and 7) knowledge presentation (visualising and presenting results to users).

## 2.2 Web Mining / Web Scraping techniques

Data mining uses data from multiple sources, printed or online. The former kept the name *data mining* (mainly structured data), whereas the latter is known as *web mining* (mostly unstructured and semi-structured). Web mining is a relatively young research area, which unsurprisingly, emerged together with the growing World Wide Web in the 1990s. Web mining was first mentioned by Etzioni (1996), who discussed the potential of mining web pages to improve people's ability to navigate, search and visualise web content. At that time, Cooley *et al*. (1997) concluded that there is "no established vocabulary" in the field of web mining and proposed a more detailed taxonomy. They differentiated between *web content mining* and *web usage mining* and offered an overview of various research issues, techniques

and development efforts. Soon after, *web structure mining* was mentioned as a third element of web mining (Madria *et al*., 1999; and Borges and Levene, 2000). Whereas *web content mining* relates to the sourcing of useful information, including text, audio and visual data; *web usage mining* refers to the investigation of user access patterns from web pages and servers. Finally, *web structure mining* describes the analysis of underlying link structures in the web (structure topology). Similar to Han *et al*. (2011), Kosala and Blockeel (2000), Zhang and Segall (2008) and Singh and Singh (2010), indicated that web mining can be separated into the following tasks:

1. *Resource finding:* The task of retrieving the intended web documents;
2. *Information selection and pre-processing:* Automatically selecting and pre-processing specific information from retrieved web resources;
3. *Generalisation:* Automatically discovering general patterns at individual web sites, as well as across multiple sites;
4. *Analysis:* Validating and/or interpreting the mined patterns (using accuracy measures);
5. *Visualisation:* Presenting the results of an interactive analysis in a visual, easy to understand fashion.

Web mining has been applied in many research areas and disciplines to extract and infer useful information from web data. Agarwal *et al*. (2008) discussed the challenge of identifying influential web bloggers. They presented a preliminary model and an approach for a robust model and conducted tests in a real-world scenario. Soriano *et al*. (2013) used information on web browsing activity and text mining to make user adapted advertisements. They reviewed different aspects of text mining and discussed strategies for specific business models. Hawwash and Nasraoui (2010) investigated web users' behaviour on websites to improve understanding and satisfying of their needs. They also proposed a web usage mining application, which is able to make use of dynamically added user data.

The current research is primarily focused on *web content mining*, applied for a variety of different objectives and purposes: e.g., customer relationship management (CRM), billing, product cataloguing and quality management. Fernández and Sleiman (2011) presented a method for data mining techniques to extract the information from the web. They compared various algorithms to obtain the best results possible, using a benchmark dataset. Chen (2014) proposed an application of a web data mining technique to enterprise the management of electronic commerce. His application used different web data mining procedures to improve efficiency in customer relationship management, to aid website construction and provide e-commerce decision support. Customer relationship management was also recently discussed in a book by Hippner and Wilde (2017), as a major area that can benefit from the advancement of web retrieval and analysis of data. There are also several reviews of various web content mining tools (as well as structure and usage) currently available: Sharma and Gupta (2012), Johnson and Gupta (2012) and Saini and Pandey (2015).

More recent work documented the role of web mining in social sciences, where behaviours and opinions can be extracted from the web and can be used to build knowledge, speed processes, and enhance collaboration. Social media (Sadilek *et al*., 2016) and education (Victor and Rex, 2016) are two important areas where capabilities of web mining have been demonstrated, especially because young people use Internet more frequently (for an overview of state-of-the-art techniques and applications in data mining, readers are referred to Shmueli *et al*., 2016 and Witten *et al*., 2016).

Yet, despite of an extant literature in web content mining, to the best of the authors' knowledge, there is no information available on web mining focusing on transport or logistics.

## 2.3 Features of the road freight industry in Australia

The Australian road freight-transport sector successfully competes with rail, water and airfreight industries and is attractive due to its price advantage, speed and convenience. The industry leads in the non-bulk freight market, with larger vehicles (e.g., b-doubles and b-triples) serving interstate and long routes, while smaller vehicles dominate the final stage of delivery. While the total number of businesses have slightly decreased from almost 43,000 to 41,097 between 2013 to 2015, their revenues increased from 51 to 52.8 billion (AUD) and their profits from 4.5 to 6.1 billion AUD (Whytcross, 2015; Gargano, 2014a). The industry is characterised by a high number of small operators and many owner-operators, who are both owners and drivers. While the four largest players contribute to only 15% of the overall revenue, the large number of small and middle-size businesses has to follow the pricing of the big companies. According to Whytcross (2015), the key external drivers of the industry are the road freight service price, the total merchandise volume of imports and exports, the world price of crude oil, the US Dollar – Australian Dollar exchange rate and the demand in wholesale trade. Gargano (2014a) highlights that the industry faces problems in productivity increase, with profit margins declining. Hence, skilled personnel, optimal capacity utilisation and effective cost controls become as important as an appropriate transport price policy and gaining long-term contracts to ensure operators' competitiveness. Therefore, understanding the market is one of the key success factors for optimum operation (Whytcross, 2015).

## 2.4 Shipper-carrier interactions and bid-price generation

Two types of transport assignments are mainly used in the road freight industry: contract logistics and single transport requests. This research is focusing on the latter. While most requests are assigned over the phone, facsimile, the Internet or face to face, there is a different degree of negotiation, depending on the value of the consignments. With the increase of e-commerce in the last two decades, online marketplaces such as Uship Inc. (Uship Inc., 2014), Teleroute (Wolters Kluwer Transport Services, 2015) or Truckit (Arcube Pty Ltd., 2015) have become key trading platforms for shippers and carriers, who can benefit from the system, by reaching out to a high number of other operators.

According to Shanahan (2003), shipping prices are determined by: 1) a carrier's base rate; and 2) negotiation process.

1. The base rate depends on the distance from origin to destination, weight of shipment, the commodity classification and accessory surcharges (i.e. breaking down the freight shipment at the destination).
2. The subsequent negotiation process not only refers to bargaining to obtain the lowest price between companies who contract transport (shippers) and the carriers, but also to the provision of detailed information on the shipped goods, which enables carriers to specify their costs more accurately and reduce the price accordingly. This information can include details about freight density, stack ability, or the composition of loads for mixed freight.

Moreover, the relationship/collaboration within both sides plays a significant role in the shipper-carrier interaction and the bid-price generation. For example, if the shipper aims to combine different shipments to simplify processes for the carrier, this is expected to lead to lower prices.

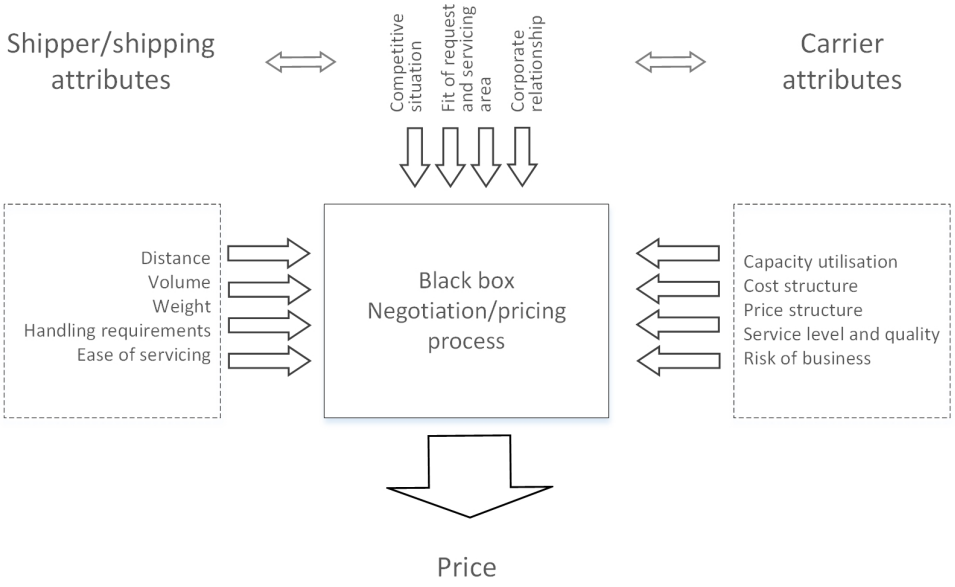
Smith *et al*. (2007)'s work confirmed the findings of Shanahan (2003). They noted the characteristics of the cargo, the number of shipments, the cargo weight, the distance, the origin and destination terminal factors, the traffic congestion and crossing of international borders as the major factors affecting the bidding price. In the overall negotiation process they also stated elements such as shippers' buying power (shipping volumes, cargo attributes, ease of servicing, risk of business) and carrier competitive position (comparative base rates, capacity utilisation, cost structure, service level and quality), as well as corporate relationships and alternative carrier options.

Although the literature in the area of bid-price generation for road freight assignments sometimes uses slightly different terms for the same elements, the findings are consistently differentiating between the two main categories, *base* and *negotiation* (for more information, the authors refer to Combes and Lafourcade, 2005; Robinson, 2013; and Marks, 2014).

Figure 1 summarises the factors that are influencing transport bid-price generation as discussed in the literature. However, detailed models are scarce and associations between the various factors are not fully explored or understood, which makes the price formation a 'black-box'.

The objective of this work is to deconstruct the pricing process by inferring the role of various attributes into the price formation through regression models. A major benefit is the potential to interrogate a website where quantitative and qualitative data on the bidding process is already available.

**Figure 1: Determinants of bid-price**




## 2.5 Rate determination models for road freight transport

Research concerned with the modelling of road-freight transport rates is not new.

Christensen and Huston (1987) examined the cost structure of specialised road freight carriers such as those with expertise in transporting building supplies, petroleum and refrigerated goods. They presented cost functions for each of these groups based on multiple regression analysis. Their findings showed, unsurprisingly, that scale economies are exploited more on major routes, and less on minor traffic routes. Thomas and Callan (1992) investigated the same groups of carriers, but their findings did not support the previous parameters,

suggesting that the aggregated data sample might have been the cause.

Swenseth and Godfrey (1996) presented different functions for freight rate estimations. Rates for 40 different routes across the US were analysed to obtain the best fitting functions. The results were compared to freight rate look-up tables and advantages in the continuous modes were detected.

In order to improve inventory replenishment and supplier selection decisions, Mendoza and Ventura (2009) estimated freight rates by calibrating two continuous rate functions, incorporated in their supplier selection model. Their experiments showed the benefits of including these elements in decision-making and the configuration of supply chains.

Özkaya et al. (2010) presented a regression model based on historical shipment rates from the USA (350,000 rates), with the objective to identify the main cost drivers in less-than-truckload (LTL) transport pricing. Their model showed a high significance and explained over 93% of the price composition. Lindsey *et al*. (2013) focused on road freight rate modelling, but also investigated the reasons for variations between rates. They presented three different regression models and found that distance, characteristics of the shipping lane and required truck type were the most important determinants.

Smith *et al*. (2007) modelled average freight rates and enabled freight forwarders to identify customers and terminals where revenues were deficient and opportunities for renegotiation arose, considering the mix of businesses. Huang *et al*. (2011) studied supply chain broker operations and their connection to possible customers. They found that not only the number of connections to customers, but also the quality of those connections increases profitability. The study of quality and quantity included the number of calls to the customer, as well as the duration a broker was associated with the company s/he was working for.

More recently, Combes (2013) presented an equilibrium price model to investigate the relationship between shipment size and tariffs, with the objective to better understand the costs structure of freight carriers. Their experiments indicated that: a) freight size close to zero does not result in a tariff close to zero; and b) the price function increases in a convex manner and tends to be constant at size levels near the capacity of vehicles.

Lately, Budak *et al*. (2017) offered two approaches to forecast truckload spot market prices. They compared an artificial neural network (ANN) and a quantile regression model and found that forecasting can be more accurate on a route-based level than when the whole network is included, which highlights the spatial aspects of the shipment. Considering these, they found the ANN model to be superior for route-based assessments, while the quantile regression model shows better results at the network level.

# 3. Data and methodology

A short Java program (combined with a couple of open source libraries) was created to extract the incoming transport requests from the website. The program ran every 15 minutes, checking requests that were *'open'*, until they were *'closed'*. Text files were then content-analysed and coded for analysis. Data on dates, quantities, locations and bids were filtered and then aggregated. A cluster analysis, followed by regression type models, identified three main categories of shipment and distinguished the most significant effects on the final bid.

Twelve variables, including both tangible (e.g., origin and destinations, distance, weight, timing, loading-unloading conditions) and intangible (e.g., flexibility, competitive pressure measured as number of 'negotiations'/bids) were considered in the pricing models.

## 3.1 Data source and extraction

Truckit.net is an online marketplace similar to Uship and Wolters Kluwer Transport Services, for trading freight transport requests within Australia. The platform presents shippers with the opportunity to upload request details, on which transporters bid. While the bidding process is structured as a reverse auction, the shipper can choose the most favourable bid, considering price, rating, transport execution time etc.

For the analysis, a web content mining application was implemented in Java language. The sequence of operations is presented in Figure A-1. It consists of two main methods: 1) identification of new transport requests; and 2) update of existing transport requests.

For the identification of new requests, the program scans the website http://www.truckit.net every 15 minutes and filters out palletised transport requests, which have not been recorded yet. The filtered requests are then downloaded into a database. Each data record includes the following: collection and delivery addresses (city, state, postcode) and dates; estimated driving distance and time; customer ID; date listed; number of bids; type of loading/unloading locations; specification of goods; shipment size (number of pallets and their size); expiry date and time of the auction, the lowest quote (AUD); status (open/closed); current last bid (in AUD); and if the bid was accepted or not. Additionally, all shipper–transporter conversations on the website's discussion forum are stored in a text file. These files often represent ungrammatical text, with shortened words, and incomplete sentences.

For the update of the existing transport requests, every eight hours the program checks every request sent to see whether there is new or additional information available. If a transport request remains unchanged, its status is changed to "closed" and no further updates are required. Likewise, the text file is updated. Rather than overwriting the information, the file is extended every time with the new data. The interval was chosen as a trade-off between duration and additional information obtained, and fine-tuned, based on observations of the frequency of changes.

## 3.3 Mixed methods

To take advantage of the richness of the secondary data and to gain from both the breadth and depth of the data from various sources, this research applied the mixed methods research (MMR) approach (Creswell and Plano Clark, 2011) in logistics. MMR reflects a unified view of research, collecting, analysing and interpreting both qualitative *and* quantitative data in a single study (or series of studies) that explores the same underlying phenomena (Leech and Onwuegbuzie, 2009; Wolf, 2010). Our method can be described as a partially mixed concurrent equal weight design, because it has: *"two phases that occur concurrently"* and *"the quantitative and qualitative phases have approximately equal weight."* (Leech and Onwuegbuzie, 2009: 268).

A combination of content analysis and multivariate data analysis was applied in keeping with our mixed methods approach. Leximancer and SPSS were the software platforms used for analysis.

### *3.3.1 Content analysis*

Recent advances in computer-aided text analysis have opened new opportunities for concept or semantic mapping. Here we applied Leximancer text analysis software for the discovery and mapping of concepts in web discussion forums (Smith and Humphreys, 2006). Leximancer provides a platform for qualitative interpretation of the communication between

customers and transporters on www.truckit.net. Within Leximancer, concepts are developed and linked through a systematic examination of the proximity with which words appear in the text. Concepts are the most semantically significant words, which are identified through frequency. The analysis is conducted by producing a co-occurrence matrix of words 'travelling together', and then a thesaurus including a list of closely related words associated by proximity to a particular concept.

A map is then built, reflecting the connections between concepts (the higher the frequency of a concept co-ocurring with another, the stronger the link between the concepts is and shorter the distance). Closer concepts are clustered in themes (see descriptions in Scott and Smith, 2005; Martin and Rice, 2007; Stockwell *et al*., 2009; Cretchley *et al*., 2010 and Caspersz and Olaru 2013, 2015). The maps are colour coded, with the more frequently occurring concepts/themes being in hot colours (red and orange) and cool colours (blue, green) and lighter shades representing the least relevant. Bigger circles indicate higher relative importance of the theme and circles with more descriptors (concepts) indicate more complex themes. The concepts 'explain' the theme and can be linked by grey lines to develop 'pathways' or connections between concepts.

Leximancer also generates transcripts for each concept and descriptors, which illustrate the story of the concept/theme, and the relationship between descriptors, when these are linked in a pathway. Finally, Leximancer has the capability to distinguish and analyse multiple sets of texts, providing an Insight Dashboard with conditional probabilities of concepts appearing within a specific set (further descriptions are available in Smith and Humphreys, 2006 and Martin and Rice, 2007).

#### *3.3.2 Quantitative analysis*

Given the exposure of the audience to the multivariate techniques and the relatively wide use of these techniques in transport modelling, a detailed description is not necessary. Instead, we clarify that clustering techniques were applied to identify patterns for freight transport. For each group, regression analysis assisted us to understand which factors determined the final bid and suggest costs for palletised transport.

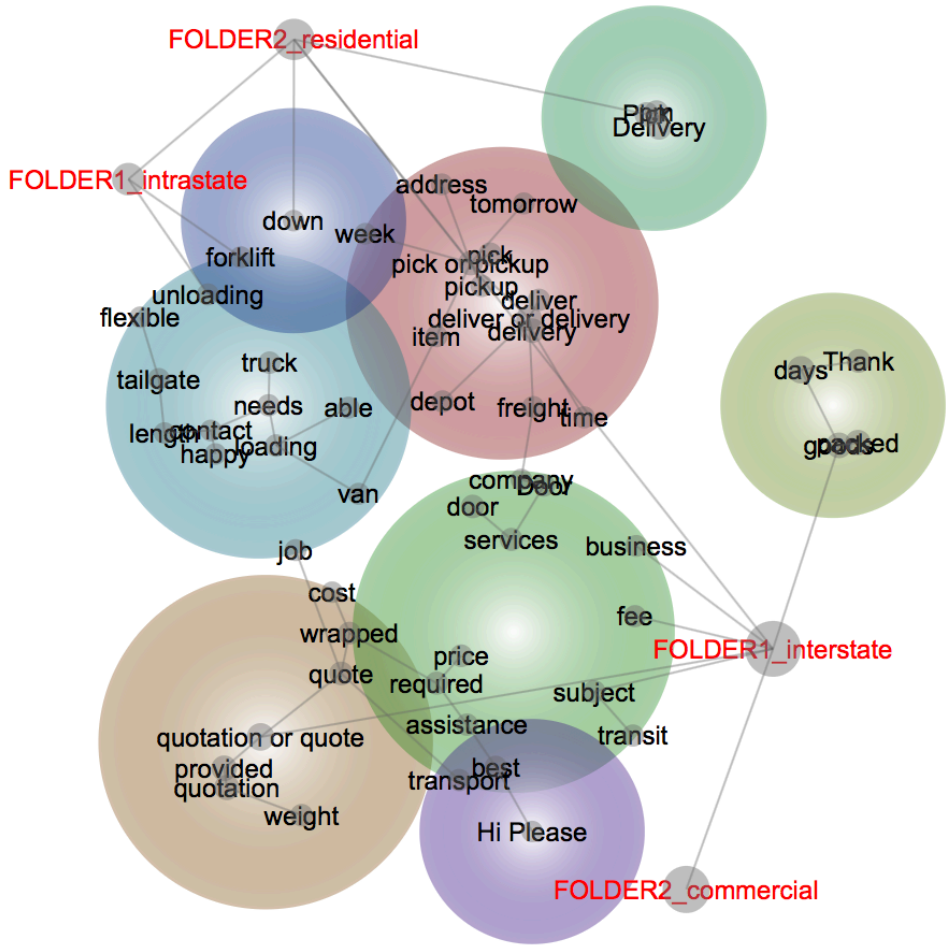
# 4. Results

Correspondence from 2,970 requests opened on the www.truckit.net website between February and December 2015 were analysed.

## 4.1 Qualitative analysis

The total of 29,737 text segments (short sentences and groups of words) were differentiated in four categories by two factors: i) interstate vs intra-state, and ii) commercial vs residential. The rank ordered Concept Lists, Thesauri, Conceptual maps and Insight Dashboards guided our interpretation. Leximancer placed each of the four groups in distinct semantic regions of the concept map, highlighting semantic differences between type of request and spatial range (Figure 2). Similarities/commonalities are evident, yet the four categories reflected these themes/concepts in various ways. To limit their influence in the Leximancer analysis, we attempted to conduct checks for inconsistent spelling and punctuation. Although some errors were found, no adjustments were made to the text excerpts, considering they were minor.

**Figure 2: Conceptual map with four categories**

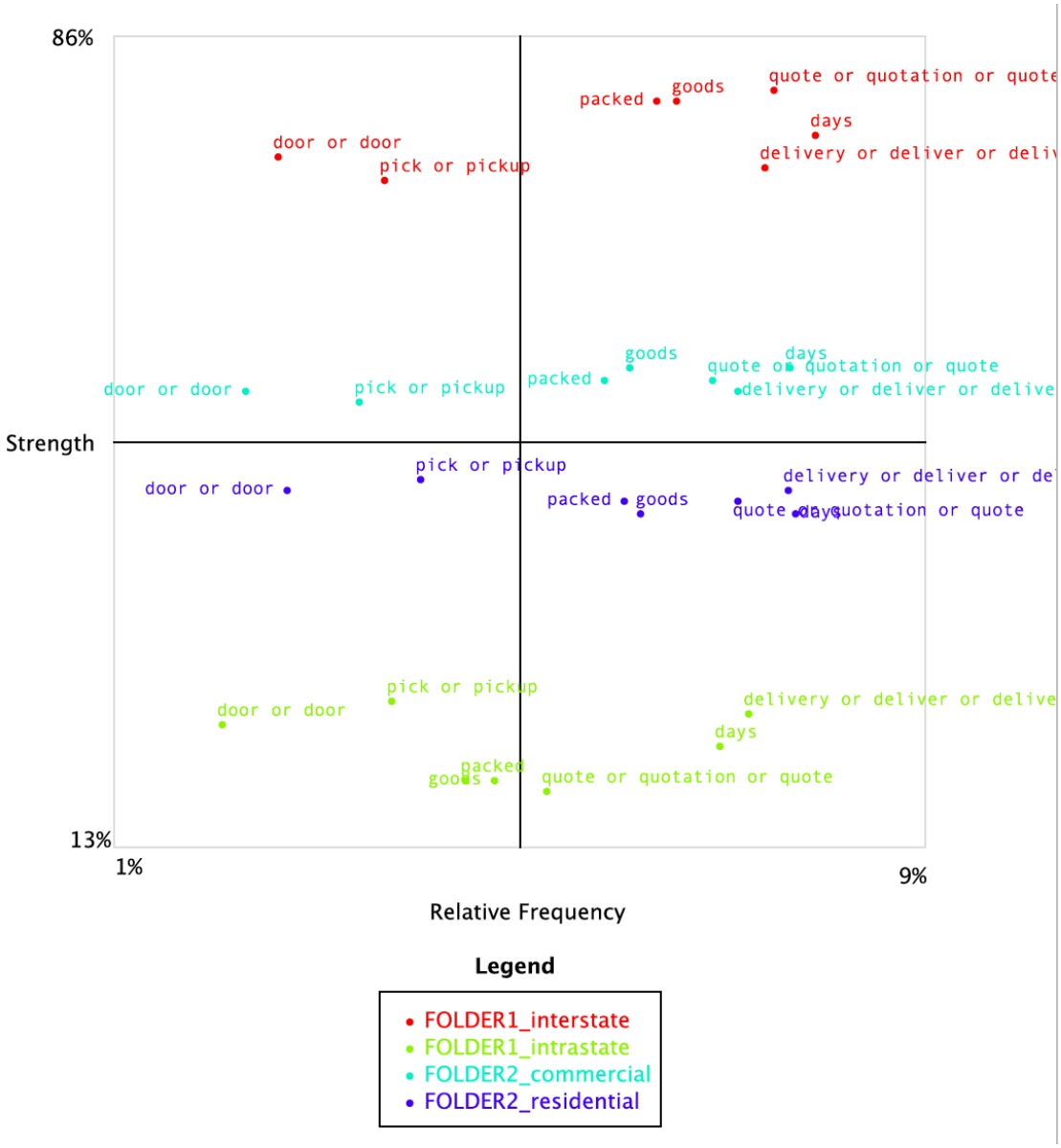



At the core of Figure 2 lies the *Delivery* theme, with the highest concentration of concepts. This is followed by *quotation, price*, and *services*. The relative closeness of the residential folder to the intrastate and commercial to the interstate confirms the spatial profile of the shipper requirements. Notably, whereas the residential requirements are closely connected to downloading, and pick-up-delivery (top-right corner of Figure 2), the courteous greetings in the message are more frequent in the commercial and interstate requests.

Figure 3 presents the strengths and relative frequency for the main concepts in the map. The clear delineation of the four sets is remarkable. While the relative frequency is similar across the four types of requests, their strength is very different. Interstate transport records strength measures above 72% for all concepts (quote, goods, packed, door to door delivery) and intrastate the lowest (<=26). The residential and commercial requests lie in the middle with strength values alike, although larger for commercial requests.

Figure 4 provides the numerical values associated with the quadrant diagram in Figure 3. The results reinforce the map, with prominence being similar for interstate, commercial, and residential requests. By contrast, focus of the intrastate requests is on time and door-to-door delivery, compared to price or packaging.

**Figure 3: Insight Dashboard of the four categories**




This output provides prominence scores, which jointly consider the strength and relative frequency (conditional probabilities). The scores show that ‘quote’, ‘packed’, ‘goods’, ‘days’, ‘delivery’ or ‘pickup’, and ‘door-to-door’ are equally and most important in the interstate shipping in Australia (top left table, in red), whereas for the intrastate transport, the ‘quote’, ‘packed’, and ‘goods’ are less critical (top right, in blue).

Again, differences can be noted between commercial and residential shippers, where timing (‘days’) represents the number one concern for companies (bottom left table), but is in the last place for residential beneficiaries (bottom right table). Conditions of shipping are similarly important, but the ‘pickup’ is key for residences.

**Figure 4: Strength and Relative Frequency for the four types of requests**

**Category: FOLDER1_interstate**

| Concept | Rel Freq (%) | Strength (%) | Prominence |
|---|---|---|---|
| quote or quotation or quote | 7 | 81 | 1.1 |
| packed | 5 | 80 | 1.1 |
| goods | 5 | 80 | 1.1 |
| days | 7 | 77 | 1.1 |
| door or door | 2 | 75 | 1.0 |
| delivery or deliver or delivery | 7 | 74 | 1.0 |
| pick or pickup | 3 | 73 | 1.0 |

**Category: FOLDER1_intrastate**

| Concept | Rel Freq (%) | Strength (%) | Prominence |
|---|---|---|---|
| pick or pickup | 3 | 26 | 1.0 |
| delivery or deliver or delivery | 6 | 25 | 1.0 |
| door or door | 2 | 24 | 0.9 |
| days | 6 | 22 | 0.9 |
| goods | 4 | 19 | 0.7 |
| packed | 3 | 19 | 0.7 |
| quote or quotation or quote | 4 | 18 | 0.7 |

**Category: FOLDER2_commercial**

| Concept | Rel Freq (%) | Strength (%) | Prominence |
|---|---|---|---|
| days | 7 | 56 | 1.0 |
| goods | 5 | 56 | 1.0 |
| packed | 5 | 55 | 1.0 |
| quote or quotation or quote | 6 | 55 | 1.0 |
| door or door | 2 | 54 | 1.0 |
| delivery or deliver or delivery | 6 | 54 | 1.0 |
| pick or pickup | 3 | 53 | 0.9 |

**Category: FOLDER2_residential**

| Concept | Rel Freq (%) | Strength (%) | Prominence |
|---|---|---|---|
| pick or pickup | 3 | 46 | 1.1 |
| delivery or deliver or delivery | 7 | 45 | 1.1 |
| door or door | 2 | 45 | 1.1 |
| quote or quotation or quote | 6 | 44 | 1.0 |
| packed | 5 | 44 | 1.0 |
| goods | 5 | 43 | 1.0 |
| days | 7 | 43 | 1.0 |

## 4.2 Quantitative analysis

### *4.2.1 Descriptive statistics*

Results indicate a substantial variability in the requests collected from the website. The shipments are dominated by commercial companies (only 14% residential, farm, or storage facilities) and vary in size with a range from 1 to 800 pallets (average of 3.83). Most requests come from two states, NSW and Victoria, and the main destinations are Qld and NSW (Table 1). There is no 'dominating' operator. On average, 2.61 bids are submitted for a request and it takes nine days to 'close' the request (Table 2).

**Table 1: O-D relationships for road transport requests**

| | | Destination State | | | | | | | | Total |
|---|---|---|---|---|---|---|---|---|---|---|
| | | ACT | NSW | NT | Qld | SA | Tas | Vic | WA | |
| Origin State | ACT | 1 | 9 | 1 | 9 | 2 | 3 | 5 | 1 | 31 |
| | NSW | 28 | **277** | 16 | **257** | 64 | 21 | 171 | 105 | 939 |
| | NT | 0 | 15 | 3 | 9 | 4 | 3 | 6 | 8 | 48 |
| | Qld | 15 | **184** | 33 | **215** | 36 | 11 | 102 | 41 | 637 |
| | SA | 3 | 54 | 7 | 39 | 17 | 4 | 38 | 36 | 198 |
| | Tas | 0 | 2 | 1 | 6 | 2 | 4 | 8 | 5 | 28 |
| | Vic | 25 | **214** | 23 | **233** | 71 | 21 | 112 | 99 | 798 |
| | WA | 3 | 57 | 12 | 52 | 20 | 6 | **70** | **71** | 291 |
| Total | | 75 | 812 | 96 | 820 | 216 | 73 | 512 | 366 | 2,970 |

Note: Abbreviations for states are as follows ACT = Australian Capital Territory; NSW = New South Wales; NT = Northern Territory; Qld = Queensland; SA = South Australia; Tas = Tasmania; Vic = Victoria; WA = Western Australia

Table 2 provides several summary statistics for the requests. Most of the requests are for transport interstate (76%), at an average distance of 1,614 km. The estimated average commercial speed is 91 km/h. In terms of urgency, only 8% of the requests have firm departure and arrival times. The bids vary accordingly from AUD 75 to AUD 29,000, with an average of AUD 787.

The data suggests substantial standardisation with more than 2/3 of the shipments using standard pallets. This is highly associated with the type of customer and 71% of the requests are downloaded at commercial establishments. Flexibility is also linked to the type of customer, and residential customers generally have fewer restrictions compared to corporate customers. The number of bids indicate that most requests do not require substantial adjustments to the price or conditions, with a reduced number involving considerable negotiation between parties.

**Table 2: Descriptive statistics for road transport requests**

| | Average (Std. deviation) or Proportion |
|---|---|
| Interstate | 76% |
| Flexibility | 8% Urgent, 25% with some flexibility, 68% Fully Flexible |
| Distance (km) | 1,613.84 (1,269.91) |
| Estimated time (h) | 17.81 (13.82) |
| Average # of bids | 2.61 (2.17) |
| | 75% up to 3 bids, but 38 requests over 10 bids |
| Unloading at residential locations | 29% of the requests |
| Use standardised pallets | 70% of the requests |
| Average Number of pallets | 3.83 (23.23) |
| Last bid (AUD) | 787 (1,343.36) |
| Closed | 94% |
| Days to close | 8.86 (8.59) |

### *4.2.2 Cluster analysis results*

Clustering refers to identifying patterns and separating objects in homogeneous groups. Each cluster has objects that are similar to others within the same cluster and dissimilar from other

clusters (Hair *et al*., 2010). Here, a multivariate clustering analysis was applied to identify patterns for freight transport and costs for pallet transport. For each cluster, multiple linear regressions then identified the factors explaining the final bid.

Twelve variables of various types (binary – e.g., flexibility of the timing; count data – e.g., number of pallets or number of bids; continuous data – e.g., bid, distance, transport duration, weight) were included in the two-stage clustering (hierarchical clustering followed by k-means clustering), using Euclidean distance on standardised data. The hierarchical, explorative stage (using an agglomerative method and comparing various algorithms – single linkage, centroid, Ward) provided a comprehensive portrayal of the potential solutions and led to the decision on three clusters as most appropriate. The seeds obtained in the hierarchical stage were used as centroids in k-means.

Three clusters emerged: 1) small interstate consignments of standardised pallets, with high flexibility; 2) special large shipments over long distances, with fixed dates for delivery and high bidding prices; and 3) flexible requests, at shorter distances and with moderate prices. The last group has the highest number of bids, and most negotiation occurs here. They also may require special (un)-loading equipment (Table 3). All multivariate tests (Pillai's trace, Wilks' lambda, Hotelling's trace, and Roy's largest root) indicate significant differences across the clusters ($p < 0.001$).

**Table 3: Comparison of the three clusters**

| Variable | 1<br>small consignments, high flexibility | 2<br>special large shipments at long distances | 3<br>flexible requests, shorter distances | Significance level (p) |
|---|---|---|---|---|
| Interstate (%) | **99** | 83 | *68* | <0.001 |
| Flexibility | **1.73** (0.538) | *1.52* (0.651) | 1.56 (0.655) | <0.001 |
| Distance (km) | **3,487.92** (700.872) | 2,121.91 (1,171.87) | *935.94* (575.167) | <0.001 |
| Estimated time (h) | **37.70** (7.819) | 23.59 (12.866) | *10.58* (6.858) | <0.001 |
| # of bids | *1.94* (1.575) | **2.83** (2.742) | 2.82 (2.260) | <0.001 |
| Unloading at residential locations (%) | **31** | *16* | 30 | <0.001 |
| Use standardised pallets (%) | **75** | 67 | 69 | 0.012 |
| Number pallets | *1.76* (2.918) | **16.93** (29.596) | 3.60 (26.231) | <0.001 |
| Last bid (AUD) | 736.74 (545.08) | **4,996.96** (3,512.23) | *496.15* (414.94) | <0.001 |
| Closed (%) | 92 | *89* | **96** | 0.011 |
| Days to close | 8.70 (8.26) | *8.51* (8.13) | **8.96** (8.70) | 0.686 |
| N | 706 | 152 | 2,081 | |

Note: In **bold** the highest values and in *italics* the lowest.

Cluster 1 includes large shipments at long distances (average 3,490 km), with flexible pick-up and delivery dates. They include the highest percentage of standardised pallets and the lowest number of bids.

Cluster 2 is the smallest and includes urgent requests with fixed dates for delivery and highest bidding prices (AUD 4,997). They usually require special loading and unloading equipment; therefore the beneficiaries are mostly commercial companies. The requests from this cluster are the largest in size (16.93 pallets on average) and go through the highest number of bids (2.83).

Cluster 3 is the largest and covers flexible requests for small intra and interstate consignments (3-4 pallets) of higher value items. The distances are the shortest (average of 936 km) and the bids are the lowest (on average AUD 496). This cluster also includes a high number of non-commercial customers.

With the exception of days to close, all other features of the requests are significantly different across the three clusters.

### *4.2.3 Regression analysis results*

A global model (with the same parameter estimates) and three different models were tested for the three clusters. Across all sub-samples, the size of the consignment (number of pallets), the distance/duration, the type of company, and flexibility are the most significant explanatory variables of the final bidding price.

Table 4 shows that each additional hour of transport is translated in AUD 12 (or 15 cents/km) and each additional pallet in AUD 7.91, everything else being kept constant. Less flexible requests cost more (on average 53.97 AUD for each level of flexibility).

Whereas there are no significant differences between clusters 1 and 3 in terms of price, a request from the cluster 2 costs AUD 4,224 more than the others.

**Table 4: Factors associated with the last bid ($R^2$-adj =0.564, se = 886.95)**

| Variables | Unstandardised Coefficients | t | Sig. |
|---|---|---|---|
| (Constant) | 441.5 | 8.881 | 0 |
| Flexibility | -53.969 | -2.065 | 0.039 |
| Estimated Time | 12.031 | 5.491 | <0.001 |
| Unloading at residential locations | -117.778 | -2.626 | 0.009 |
| # of Pallets | 7.908 | 11.138 | <0.001 |
| Cluster 1 small consignments, high flexibility | -49.343 | -0.699 | 0.485 |
| Cluster 2 special large shipments at long distances | 4,224.367 | 52.26 | <0.001 |

The separate analyses show that only two characteristics are significant across all clusters: distance and size (Table 5). This is important, as it shows relative stability of the 'cost structure' in the past decade. For example, previous work reported by Smith *et al*. (2007), highlighted the number of shipments, their weight and distance as cost drivers for the road freight industry. However, flexibility in terms of timing affects the bidding price for cluster 3 requests and type of customer for the bids in clusters 1 and 3. The fit measures are poorer and likely affected by the influential outliers in each cluster.

**Table 5: Factors associated with the last bid by cluster**

| | Cluster 1 small consignments, high flexibility | | | Cluster 2 special large shipments at long distances | | | Cluster 3 flexible requests, shorter distances | | |
|---|---|---|---|---|---|---|---|---|---|
| **Variables** | **Unstd. Coeff.** | **t** | **Sig.** | **Unstd. Coeff.** | **t** | **Sig.** | **Unstd. Coeff.** | **t** | **Sig.** |
| (Constant) | 515.73 | 4.318 | <0.001 | 2,679.89 | 3.595 | <0.001 | 479.082 | 18.259 | <0.001 |
| *Flexibility* | *-13.609* | *-0.374* | *0.709* | *-388.644* | *-1.084* | *0.280* | -38.892 | -2.85 | 0.004 |
| **Distance (km)** | **0.051** | **1.816** | **0.070** | **0.803** | **3.920** | **<0.001** | **0.094** | **6.037** | **<0.001** |
| Unloading at residential locations | -104.943 | -2.309 | 0.021 | *-235.367* | *-0.197* | *0.844* | -128.743 | -5.557 | <0.001 |
| **# of Pallets** | **53.908** | **7.879** | **<0.001** | **72.073** | **8.871** | **<0.001** | **2.361** | **6.972** | **<0.001** |
| **GOF** | ($R^2$-adj =0.096, se = 519.67) | | | ($R^2$-adj =0.342, se = 2,848.37) | | | ($R^2$-adj =0.490, se = 404.10) | | |

Note: Diagnostic tools, such as residual plots are not included in the paper, as they are not deemed essential for the objective of the research. In this case there were no violations of the LINE/BLUE assumptions. Outliers were identified (Mahalanobis distances) and the analysis was run with and without them to check their impact. All VIFs were below 5.

Two examples of prediction using the models can show the price formation:

1) We assume a shipment of 10 pallets of wine from the South-West of WA to the metropolitan area of Perth. The average distance is 400 km and the expected time is 10 hours (cluster 3 type request).

   By applying the full sample model we obtain: 441.5 – 2 * 53.969 +10 * 12.031 + 10 * 7.908 = AUD 532.95 (CI – 494.23; 571.62).

   By applying the cluster 3 model we have: 479.08 – 2 * 38.892 +400 * 0.094 + 10 * 2.361 = AUD 462.51 (CI – 444.70; 480.32).

2) The second example considers 50 pallets of brick shipped interstate between NSW and Qld to a residential site. The average distance is 700 km and the expected time is 48 hours (cluster 2 type request).

   By applying the full sample model we obtain: 441.5 – 2 * 53.969 + 48 * 12.031 – 1 * 117.781 + 50 * 7.908 + 1 * 4,224.37 = AUD 5,413.04 (CI – 5,005; 5,822).

   By applying the cluster 2 model we have: 2,679.89 – 2 * 388.644 + 700 * 0.803 - 1 * 235.367 + 50 * 72.073 = AUD 5,832.98 (CI – 5,359.62; 6,306.34).

Although the values obtained with the two models are not the same, they are more similar to each other and substantially different from the averages that would have been obtained using as predictors only the number of pallets and distance (AUD 13.44 per additional pallet; AUD 0.184 per km). The 'traditional' approach would overestimate small consignments (AUD 647.14 vs 450-500 in the wine example) and heavily underestimate the larger requests (AUD 1,239.95 vs 5,000-6,000 for the bricks). This is a useful demonstration of the benefits of applying more detailed models, especially when the data can be obtained from secondary sources.

# 5. Discussion and conclusion

In Australia, small operators, contributing 85% to the overall market revenue, dominate the road freight industry. This means that specific approaches are required to model the particular 'atomic' structure of the market.

To determine the influences on bid-price generation, a number of common factors that describe transport requests on online marketplaces have been mined, using data from www.truckit.net. A short Java program was created to extract the incoming requests. The combination of content analysis and quantitative analysis enabled a richer understanding of the freight market operation.

There are substantial differences between interstate and intrastate shipments, as well as between commercial and residential customers. For many residential and or intrastate transports, customers are primarily concerned with unloading settings and flexibility of times; whereas for transport at longer distances, interstate and between commercial companies, the door-to-door conditions and the prices/quotes appear more frequently in their communication.

The quantitative analysis revealed three broad types of requests: interstate, flexible average-size shipments (a quarter of the total number of requests); special large transport tasks (5-6%); and intrastate short-distance shipments (around 70%). Unsurprisingly, distance and number of pallets affect the bidding price, but this is also impacted by unloading conditions and flexibility of terms (for intrastate and residential shipments). This is a common finding from both qualitative and quantitative analyses.

Although traditionally companies use transport distance, load volume and weight for setting their bids, many other factors (type of customer, distance/time, flexibility) could be easily incorporated in the quotes, to better differentiate various types of shipments and consequently increase their chance of winning profitable bids. The prediction examples offered show that the common practice is likely to overestimate prices for small shipments and heavily underestimate large shipments.

In conclusion, the mixed-methods analysis revealed the benefits of extracting and analysing secondary data from readily available sources. This is particularly relevant in cases where primary data collection is difficult. Understanding the potential internal mechanisms used by small operators in organising their operation and determining their prices assists us in designing (and validating with real data) a decision support RM system for application in road freight transport. The results of the research could inform industry/companies about prices and spare capacity, which can then act towards increasing their revenues and efficiency/productivity. As the approach evaluates the whole price structure and possible revenues, companies would be able to choose their most profitable requests.

Future research could explore the possibility to combine secondary data with some short surveys conducted by the companies using the website. Additionally, non-linear models, with or without interactions, may be tested for larger samples of data.

# 6. References

Agarwal, N, Liu, H, Tang, L, Yu, PS, 2008, Identifying the influential bloggers in a community, *Proceedings of the 2008 International Conference on Web Search and Data Mining*. ACM, Palo Alto, California, USA, pp. 207-218.

Arcube Pty Ltd, 2015, TRUCKIT.NET, *About Us*, available at https://www.truckit.net/about-us,

Borges, J, Levene, M, 2000, Data Mining of User Navigation Patterns, In: Masand, B., Spiliopoulou, M. (Eds.), *Web Usage Analysis and User Profiling: International WEBKDD'99 Workshop San Diego*, CA, USA, August 15, 1999 Revised Papers. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 92-112.

Budak, A, Ustundag, A, Guloglu, B, 2017, A forecasting approach for truckload spot market pricing. *Transportation Research A* 97, pp. 55-68.

Bureau of Infrastructure Transport and Regional Economics (BITRE), 2014, *Freightline I - Australian freight transport overview*. BITRE, Canberra.

Bureau of Transport and Regional Economics, 2003, *Working paper 60 - An overview of the Australian road freight industry*, Canberra.

Caspersz, D, Olaru, D, 2013, Developing 'emancipatory' interest: learning to create social change, *Higher Education Research and Development (HERDSA) Journal*, 33(2), pp. 226-241.

Caspersz, D, Olaru, D, 2015, The Value of Service learning: The Student Perspective, *Studies in Higher Education*, DOI: 10.1080/03075079.2015.1070818.

Chen, G, 2014, Application of Web Data Mining Technique to Enterprise Management of Electronic Commerce, *Seventh International Symposium on Computational Intelligence and Design* (ISCID). IEEE, pp. 154-157.

Christensen, LR, Huston, JH, 1987, A Reexamination Of The Cost Structure For Specialized Motor Carriers. *Logistics and Transportation Review* 23(4), p.339.

Combes, F, 2013, On Shipment Size and Freight Tariffs Technical Constraints and Equilibrium Prices. *Journal of Transport Economics and Policy* 47(2), pp. 229-243.

Combes, P-P, Lafourcade, M, 2005. Transport costs: measures, determinants, and regional policy implications for France. *Journal of Economic Geography* 5(3), pp. 319-349.

Cooley, R, Mobasher, B, Srivastava, J, 1997, Web mining: information and pattern discovery on the World Wide Web, *Proceedings of the Ninth International Conference on Tools with Artificial Intelligence*. IEEE, pp. 558-567.

Creswell, JW, Clark, VLP, 2011, *Designing and Conducting Mixed Methods Research*, 2 ed. SAGE Publications, Los Angeles.

Cretchley, J, Gallois, C, Chenery, H, Smith, A, 2010, Conversations Between Carers and People With Schizophrenia: A Qualitative Analysis Using Leximancer. *Qualitative Health Research* 20(12), pp. 1611-1628.

Etzioni, O, 1996. The World-Wide Web: quagmire or gold mine? *Communications of the Association for Computing Machinery* 39(11), pp. 65-68.

Fernández, G., Sleiman, H., 2011, An Experiment on Using Datamining Techniques to Extract Information from the Web, In: Corchado, J., Pérez, J., Hallenborg, K., Golinska, P., Corchuelo, R. (Eds.), *Trends in Practical Applications of Agents and Multiagent Systems*. Springer Berlin Heidelberg, pp. 169-176.

Gargano, S, 2014a, *IBISWorld Industry Report I4610 - Road Freight Transport in Australia*. Available at: http://www.ibisworld.com.au.

Gargano, S, 2014b, *IBISWorld Industry Report I5292A - Road Freight Forwarding in Australia*. Available at: http://www.ibisworld.com.au.

Gargano, S, 2014c, *IBISWorld Industry Report X0016 - Integrated Logistics in Australia*. Available at: http://www.ibisworld.com.au.

Hair, JF, Black, WC, Babin, BJ, Anderson, RE, 2010, *Multivariate data analysis*, 7 ed. Pearson, Upper Saddle River, N.J.

Han, J, Kamber, M, Pei, J, 2011. *Data mining: concepts and techniques: concepts and techniques*. Elsevier, Boston.

Hawwash, B, Nasraoui, O, 2010. Mining and tracking evolving web user trends from large web server logs. *Statistical Analysis and Data Mining* 3(2), pp. 106-125.

Hedley, J, 2009-2015. *Jsoup*. Available at: http://jsoup.org/apidocs/org/jsoup/package-summary.html.

Hippner, H, Wilde, KD, 2017, *Data Mining im CRM, Effektives Customer Relationship Management*. Springer, pp. 141-158.

Huang, M, Homem-de-Mello, T, Smilowitz, K, Driegert, B, 2011, Supply Chain Broker Operations: Network Perspective. *Transportation Research Record* 2224, pp. 1-7.

Joda.org, 2015, *Joda-Time*. Available at: http://www.joda.org/joda-time/ (accessed 05/03/2015).

Johnson, F, Gupta, SK, 2012, Web Content Mining Techniques: A Survey. International *Journal of Computer Applications* 47(11).

jxl, 2015. J *Excel Api*, Available at: http://www.jexcelapi.sourceforge.net.

KordaMentha, 2012, *Road freight part 1 - Industry overview*. Available at: http://www.333group.com/docs/publications/12-08_road-freight-industry_part-1.

Kosala, R, Blockeel, H, 2000, Web mining research: a survey. *The Association for Computing Machinery's Special Interest Group on Knowledge Discovery and Data Mining Explorations Newsletter* 2(1), 1-15.

Leech, NL, Onwuegbuzie, AJ, 2009, A typology of mixed methods research designs. *Quality & Quantity* 43(2), pp. 265-275.

Lindsey, C, Frei, A, Ali Babai, H, Mahmassani, HS, Park, Y-W, Klabjan, D, Reed, M, Langheim, G, Keating, T, 2013, Modeling Carrier Truckload Freight Rates in Spot Markets. *92nd Annual Meeting of the Transportation Research Board*, Washington D.C.

Madria, SK, Bhowmick, SS, Ng, W-K, Lim, EP, 1999, Research Issues in Web Data Mining, In: Mohania, M., Tjoa, A.M. (Eds.), *Data Warehousing and Knowledge Discovery: First International Conference*, DaWaK'99 Florence, Italy, August 30 – September 1. Springer Berlin Heidelberg, pp. 303-312.

Marks, D, 2014, 4 of the most important factors that determine LTL freight rates. Available at: http://www.getwindfall.com/blog/4-of-the-most-important-factors-that-determine-ltl-freight-rates.

Martin, NJ, Rice, JL, 2007, Profiling Enterprise Risks in Large Computer Companies Using the Leximancer Software Tool. *Risk Management* 9(3), pp. 188-206.

Mendoza, A, Ventura, JA, 2009, Estimating freight rates in inventory replenishment and supplier selection decisions. *Logistics Research* 1(3), 185-196.

Özkaya, E, Keskinocak, P, Roshan JV, Weight, R, 2010, Estimating and benchmarking Less-than-Truckload market rates. *Transportation Research E* 46(5), pp. 667-682.

Robinson, A, 2013, *10 Factors which determine LTL freight rates*. Available at: http://cerasis.com/2013/11/19/ltl-freight-rates/.

Sadilek, A, Kautz, HA, DiPrete, L, Labus, B, Portman, E, Teitel, J, Silenzio, V, 2016, *Deploying nEmesis: Preventing Foodborne Illness by Data Mining Social Media*, AAAI. Citeseer, pp. 3982-3990.

Saini, S, Pandey, HM, 2015, Review on Web Content Mining Techniques. *International Journal of Computer Applications* 118(18), pp. 33-36.

Scott, N, Smith, AE, 2005, Use of Automated Content Analysis Techniques for Event Image Assessment. *Tourism Recreation Research* 30(2), pp. 87-91.

Shanahan, J, 2003. Demystifying LTL pricing. *Logistics Management* 42(10), pp. 31-36.

Sharma, AK, Gupta, P, 2012, Study & Analysis of Web Content Mining Tools to Improve Techniques of Web Data Mining. *International Journal of Advanced Research in Computer Engineering & Technology* (IJARCET).

Shmueli, G, Patel, NR, Bruce, PC, 2016, *Data Mining for Business Analytics: Concepts, Techniques, and Applications with XLMiner*. John Wiley & Sons.

Singh, B, Singh, HK, 2010, Web Data Mining research: A survey, *2010 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)*, pp. 1-10.

Smith, AE, Humphreys, MS, 2006, Evaluation of unsupervised semantic mapping of natural language with Leximancer concept mapping. *Behavior Research Methods* 38(2), pp. 262-279.

Smith, LD, Campbell, JF, Mundy, R, 2007, Modeling net rates for expedited freight services. *Transportation Research Part E* 43(2), pp. 192-207.

Soriano, J, Au, T, Banks, D, 2013, Text mining in computational advertising. *Statistical Analysis and Data Mining* 6(4), pp. 273-285.

Stockwell, P, Colomb, RM, Smith, AE, Wiles, J, 2009, Use of an automatic content analysis tool: A technique for seeing both local and global scope. *International Journal of Human-Computer Studies* 67(5), pp. 424-436.

Swenseth, SR, Godfrey, MR, 1996, Estimating freight rates for logistics decisions. *Journal of Business Logistics* 17(1), pp. 213-231.

Thomas, JM, Callan, SJ, 1992, Cost Analysis of Specialized Motor Carriers: An Investigation of Aggregation and Specification Bias. *Logistics and Transportation Review* 28(3), p.217.

Uship Inc., 2014, *Uship transportation network*. Available at: http://\-www.\-uship\-.com/au/.

Victor, S, Rex, MX, 2016, Analytical implementation of web structure mining using data analysis in educational domain. *International Journal of Applied Engineering Research* 11(4), pp. 2552-2556.

Whytcross, D, 2015, *IBISWorld Industry Report I4610 - Road Freight Transport in Australia*. Available at: http://www.ibisworld.com.au/industry/default.aspx?indid=456.

Witten, IH, Frank, E, Hall, MA, Pal, CJ, 2016, *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

Wolf, F, 2010, Enlightened Eclecticism or Hazardous Hotchpotch? Mixed Methods and Triangulation Strategies in Comparative Public Policy Research. *Journal of Mixed Methods Research* 4(2), pp. 144-167.

Wolters Kluwer Transport Services, 2015, Teleroute Freight Exchange. Available at: http://teleroute.com.

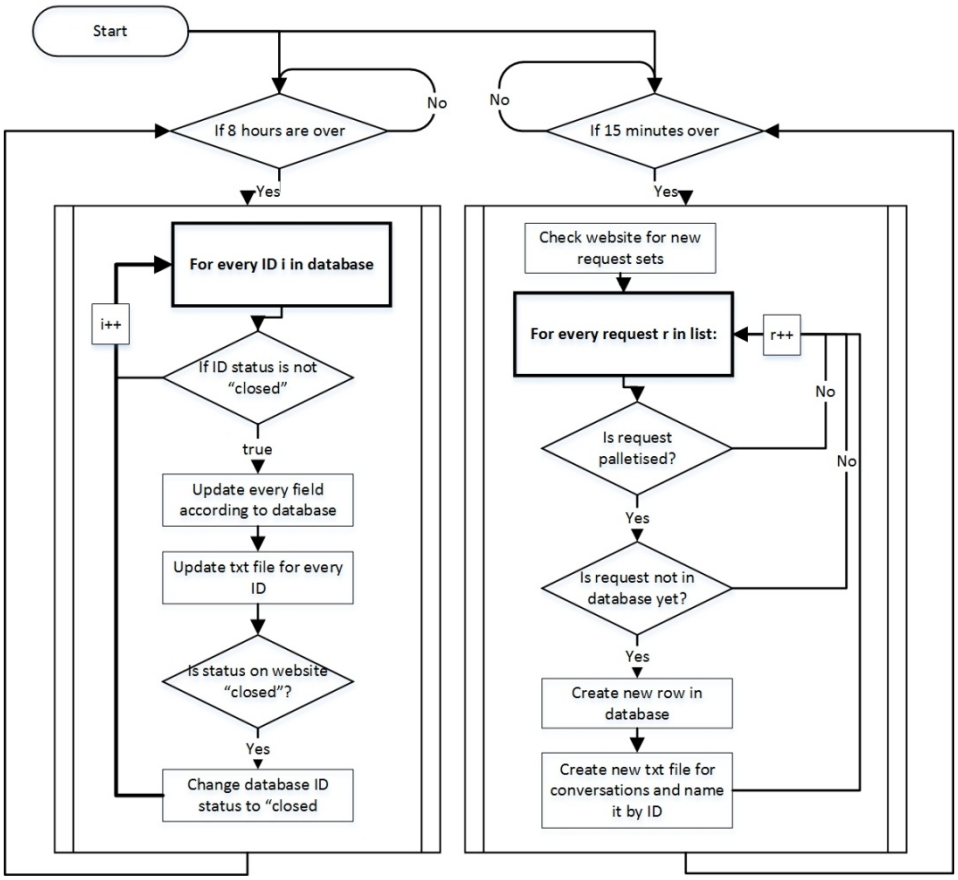
Zhang, Q, Segall, RS, 2008, Web Mining: A Survey of Current Research, Techniques, and Software. *International Journal of Information Technology & Decision Making* 07(04), pp. 683-720.

# Appendix - Implementation of data mining

The program implemented in Java language included two stages: 1) identification of new transport requests; and 2) update of existing transport requests. The sequence of operations is presented in **Error! Reference source not found.**.

The program checked every request sent, to see whether there is new or additional information available. If the information was new ("new request"), the program created a new row in the database and a new text file storing the communication between shipper and its potential carriers. The text file was updated until the request status changed to "closed". Rather than overwriting the information, the file was extended every time with the new data.

**Figure A-1: Web scraping algorithm flowchart**




For the Java implementation, the following packages were used: "Jsoup" as a Java HTML parser library (Hedley, 2009-2015), "JExcelApi" to read, write, and modify the Excel spreadsheets (jxl, 2015) and "Joda-Time", as a replacement for the Java date and time classes (Joda.org, 2015).