Delft University of Technology

FACULTY MECHANICAL, MARITIME AND MATERIALS ENGINEERING
Department Maritime and Transport Technology

Mekelweg 2
2628 CD Delft
The Netherlands
Phone +31 (0)15-2782889
Fax +31 (0)15-2781397
www.mtt.tudelft.nl

Specialization:	Transport Engineering and Logistics
Report number:	2014.TEL.7888
Title:	Assessment of raw custom's transport data based on container ratios
Author:	F.M.W. Chilla

Title (in Dutch)	Beoordelen van onbehandelde douane transport data gebruik makend van container ratios

Assignment:	research
Confidential:	no
Initiator (company):	M. Koclega (Seabury, Amsterdam)
Supervisor:	Y. Pang
Date:	23 October 2014

TU Delft
Delft University of Technology

FACULTY OF MECHANICAL, MARITIME AND MATERIALS ENGINEERING

Department of Marine and Transport Technology

Mekelweg 2
2628 CD Delft
Phone +31 (0)15-2782889
Fax +31 (0)15-2781397
www.mtt.tudelft.nl

Student:	F.M.W. Chilla	Assignment type:	Research
Supervisor (TUD):	Y. Pang (TU Delft)	Creditpoints (EC):	15
Supervisor (Company)	C. Given (Seabury)	Specialization:	TEL
		Report number:	2014.TL.7888
		Confidential:	No

Subject: **Assessment of raw custom's transport data based on container ratios**

Seabury Cargo Advisory BV builds and maintains a demand database on containerised maritime trade derived from customs data. The percentage ratio of containerisation per trade flow is determined from a small group of reporting countries whose customs data distinguishes containerised tonnage.

At the moment all of the analyses are performed in database software. However, this software does have statistical limitations and is not suitable to build a predictive model. Therefore, Seabury would like to gain insight in how to handle the raw customs data in R. Using this methodology, the student should be able to determine the accuracy of the containerisation data issued by different customs data sources. Based on this initial analysis, a new model can be derived that more accurately estimates containerisation for reporting countries that provide this data, but also for trade flows where no containerisation data exists. Hence, this study should give an intuition how the CR can be incorporated in a model and how the different data sources are behaving.

Your assignment is to:

- Develop a method to handle the raw customs data in R
- Analyse the reliability of the containerisation data for the following reporting sources: US customs, UK customs, Spain customs, Japan customs, Taiwan customs and Eurostat.
- Determine the commodities that fluctuate between container transport and non-container transport methods in more detail and determine the drivers of their containerisation.

The report should comply with the guidelines of the section. Details can be found on the website.

The mentor,

Y. Pang

Summary

In this report, an explanatory study was conducted on so-called container ratios. A container ratio (CR) denotes the distribution of commodities shipped in containers, against commodities shipped outside of a container. This metric is able to reveal underlying trends and assess the reliability of the dataset. To our knowledge, not prior research was conducted on CR's in this context. Only a hand full of international customs issue container specific information, namely the UK, US, Taiwan, Japan, Spain and an aggregate of European counties. A large amount of transport companies however, are interested in this information for countries not issuing any information as well. Before an assessment could be conducted, problems associated with the data format had to be overcome. Due to a lack of validation points, a direct comparison between the sources was conducted. Taiwan and Japan showed more consistent results compared to the other sources. This information could be used in the next phase of the project, where a model will be built estimating the CR's for the mentioned countries.

One of the main objectives of the report was to come up with a methodology that was able to handle the data size and complexity. The dataset was aggregated upon its country source and loaded in the software package R. Only the active parts of the dataset were loaded into the main memory using SQL, while keeping track of metadata of the whole object, also the non-active parts. Moreover, a number of operations are executed to reduce data complexity. A weighing factor was deployed for the issued CR's, for aggregation purposes the Split-Apply-Combine structure was used, the labelling was made uniform and the dataset was cleaned. In the end, the proposed scheme led to a fast and lean implementation of the dataset.

After the data complexity was reduced, the influence of different variables on the CR could be explored. The category of the shipped product proved to be a good proxy for the CR, both on an industry level as on a more specific level. Moreover, the historical movement of a CR was an effect that could not be neglected. When shifting to deeper categorical levels, classes with high weights had a high influence. There seemed to be little grouping possible within the regional variable, based on the CR. Next, the different datasets were compared to each other and where possible, an ordering of the sources was formed. Hence, the qualities of the different data sources were assessed. First the historical CR's were analysed for different subgroups. On a high global level, the volatility and fluctuations of the CR's were striking. When comparing the CR's of similar categories for different datasets, Taiwan and Japan denoted the most stable and reliable CR series. A second method to test the reliability of the different data sources was by comparing the importing trade flows from dataset A to dataset B, to the exporting trade flows of dataset B to dataset A. The datasets of Taiwan and Japan have shown similar movement concerning the CR's through time. Trade flows that were

reported in those two different sources, showed identical numbers. When combining these results with earlier findings, the parallels could be drawn. The Eastern countries have shown consistent and explainable results for the different industry categories.

Finally, it was researched whether the CR depended on the aggregated monthly-shipped volume for selected bulk materials for all the datasets. This illustrated how a CR could be used to detect underlying trends in trade data, which would be hard to discover using traditional trade numbers. To measure the relationship, a regression was conducted in which the significance of transported weight on the CR was tested. The US dataset has denoted a relationship where a smaller shipped weight resulted in higher CR's. This is line with expectations, since the trade imbalance forces countries to come up with new ways of dealing with empty containers. The results for the other countries were more volatile. Hence, the use of CR in this context proved its added value.

The findings of this research can be used as a start point for future research. The ranking need to be substantiated before it can be used in a model. Building a reliable model could ultimately lead to more efficient supply chain logistics, with all the economical and ecological advantages this implies.

List of Figures

Contents

1. Introduction

1.1 General introduction

With the global transport sector developing non-stop, transportation companies are growing fast and optimal logistics are getting increasingly complex. The role of containers in this process cannot be neglected (as pointed out by (Vigarie, 1999) or (Rijsenbrij & Van Ham, 2012)). Within the competitive transportation landscape, having all the relevant information at any point in time is crucial. This implies not only considering the absolute magnitude of trade flows, but also spotting the underlying drivers of these trade flows on time. Effective planning could lead low transport costs and environmental advantages.

Seabury, a maritime data supplier, has trade data coming in from different custom's offices around the world. Data is collected, controlled content-wise and made available for clients. Reporting countries of those custom's offices have monthly data supplied for each trade-lane (origin and destination country) on a commodity level (possibly for different modes of transportation). For UK, US, Spain, Taiwan and Japan only, the data has containerization level of detail. Hence, within these country sources, for every issued trade lane a separation is made: the volume shipped containerized is decoupled from the volume that is not shipped in shipping containers. This information is very interesting from a business perspective, as it can denote trends in preferred transport method for certain commodities. Several authors have discussed the necessity of researching the coordination of related containers shipments (Fransoo, 2008).

Given that some customs issue trade information on the maritime transport method (containerized or non-containerized), a useful metric can be constructed: the Container Ratio (or CR). The CR basically denotes a ratio which reflects a specific trade volume shipped by containers, divided by the total volume in that trade lane. For example, ski boots are very likely to be solely shipped containerized. The shipped volume of ski boots in containers is therefore equal to the total shipped volume, leading to a CR of one. Hence, the CR depends of the type of cargo considered. The CR helps to gain insight how materials are shipped and anticipate on that. In this report, the commodities are mainly reconnoitred on a high level.

As said, only a selection of countries issues containerized information. Furthermore, there is a group of countries, which does not issue containerized information and a group of countries

issuing no trade data at all. Realizing that this classification exists on both the importing and exporting side of trade lanes, one is able to create Figure 1. It is given that a trade flow has two sides; one importing and one exporting side. By combining the information yielded in one of those side, it is possible to construct a more complete transport matrix. For classes 1, 2 and 3 in the figure, containerized trade data is available for at least one of the involved countries. For regions 4 and 5, the trade data is available but not with a containerized level of detail. For region 6, no monthly data is accessible. The UN for example gives out yearly data, which has proven to be less reliable.

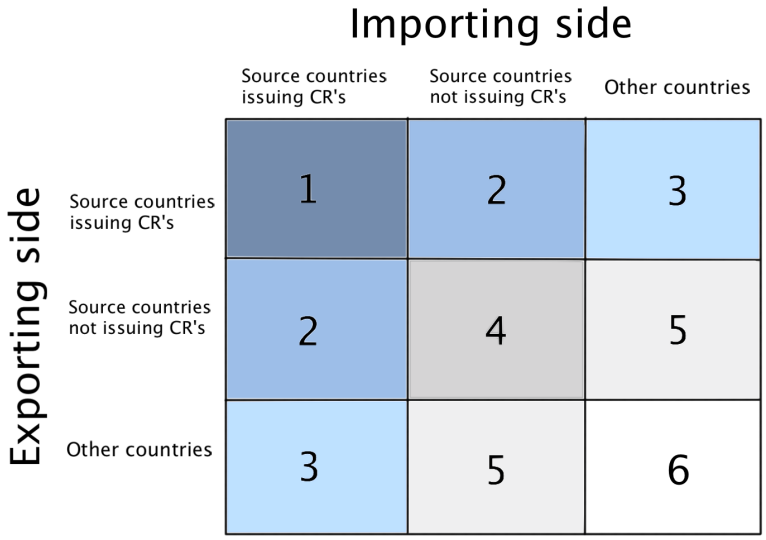


Figure 1: Classification of the information delivered by customs

As said, transportation companies are very interested in the CR of certain trade lanes or commodities. Although the amount of countries issuing this information is limited, Seabury is interested in possibilities to model this missing data based on the provided containerization data of quadrant one to three. However, before such extrapolation studies become useful, one needs to get an intuition of the general behavior of CR's through time and reliability of the different issuing countries. Hence, the data sources are analyzed individually and compared to each other. In Figure 1, this implies that our scope is only limited to quadrant one. One could see this study as a start point, from which a statistical model estimating the container information for other sources could be created in the future. Accurate forecasting of container flows is essential in minimizing overall risks and realizing significant yield improvements (Ramanakumar, 2009).

Another problem revolves around the handling of the raw customs data. Because the data is coming different sources, the consistency and structure among them is limited. This makes direct comparisons not possible without generalizing it in some way. Moreover, since all the small trade lanes are included in the original set, this research involves very large datasets. Regular RAM loading schemes are not feasible, a work around scheme should be constructed. Before researching the CR of assessing the quality of the issuing source, this data complexity needs to be handled. Hence, a method should be developed able to decrease the complexity of the data.

To our knowledge this is the first study in which CR's are used to denote underlying trends or assess the quality of datasets. However, given the prior information of containerization for some commodity classes (like the mentioned ski boots), quality assessments are feasible. Validation based on the absolute size of trade lanes is not possible, given that no references exist. Therefore, in this report the assessment is limited using the CR's. This report should be considered as a starter upon which choices for the predictive containerization model can be substantiated with balanced choices.

1.2 Goal of the Research

Given the problems described in the previous section, an objective for this report can be formulated:

The goal of this research is to assess the quality of the data sources of Seabury, using the issued containerized trade flows

In order to achieve this objective, a number of underlying questions should be answered. The research questions are:

1. Is the container ratio an appropriate metric to assess the data sources?
2. How can the complexity of the raw customs data be reduced in order to perform the analyses?
3. To what extend do the categorical and regional characteristics of the trade flows have influence on the container ratios?
4. How do the different data sources compare and is a classification possible based on the analyses?
5. Does the container ratio depend on the aggregated monthly-shipped volume for selected bulk materials?

To measure the reliability of a set of data in general, several measurements of the same units/variables should be analysed and compared. Here Seabury does not have several measurements of the same units for the same instance, but has measurements of the same units over time. First reordering and structuring the raw data should be considered.

Moreover, the datasets cannot be validated using a dataset with the "true" numbers. Hence, other measures should be used to approach the realistic container ratio. If not possible, consistency will be assumed to be a good starting point for reliability. Seabury is interested in distinguishing the commodities, for which the containerization ratio fluctuates the most, and the reasons driving such volatility on high level.

1.3 Structure of the report

The structure of the remainder of this report is as follows. In Chapter 2 background information is given. A classification of cargo is given and the urge to use the CR is explained. In Chapter 3 the methodology used to research the CR's is given. The handling of the large datasets is discussed, how the data cleaned in order to conduct analyses on them and how reliability is tested are discussed in this chapter. In Chapter 4 the data and results of the data handling are discussed together with the results of the CR analyses. The concluding remarks and a discussion can be found in Chapter 5. This structure is depicted in Figure 2.

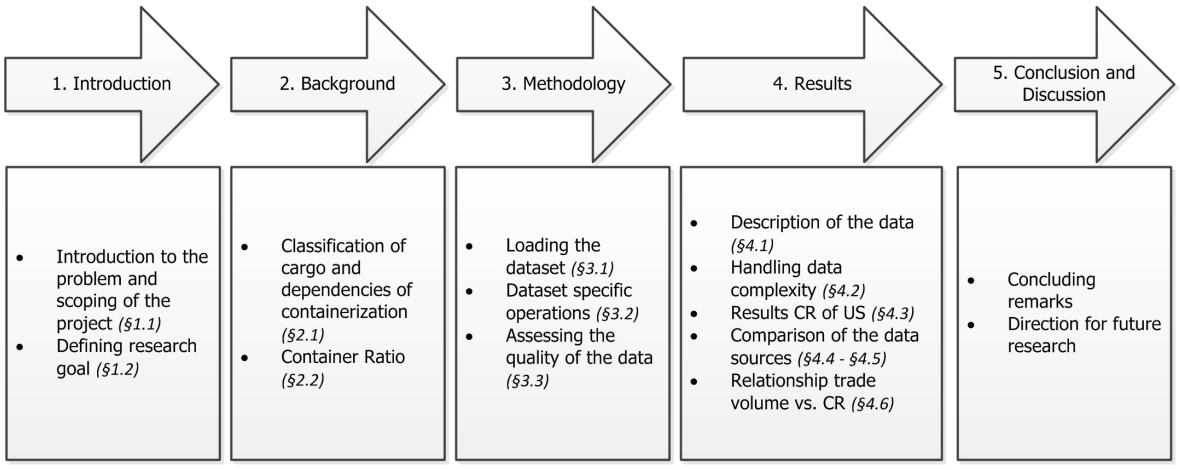


Figure 2: A diagram of the structure of the report

2. Background

In this chapter information is given concerning the context of the research scope. The goal of this chapter is inform the reader about the transport/commodity types on a high level, elaborate on containerization and give a motivation for the use of CR in the research. A general introduction into cargo with the two different types of cargo are defined and presented in Section 2.1. In Section 2.2 some of the factors that could influence the containerization are listed, together with background in which the Container Ratio (CR) metric is introduced and explained.

For several reasons it is important accurately estimating the direction and magnitude of container flows. Tavasszy (Tavasszy et al., 2011) show that for costs optimization, ports and hinterlands infrastructure need to have reliable projections of trade flows. Infrastructure can only be designed appropriately if the projected container flows are accurate. Tavasszy (Tavasszy et al., 2011) propose a model, which is able to make strategic choices for container shipping routes based on freight rates. Hence, observing trends in containers lanes on time is critical in making forecasts for port infrastructure.

Also when considering a smaller time horizon, having accurate forecasts of container streams is essential. Transportation companies can make a planning of their inventory based on their projected transport. Moreover, if a certain material is shifting from bulk transport to containerized transport, this could have consequences for the inventory and capacity of ports. And, with freight rates leading when it comes to transport types and port choices (Grossmann, 2007), having information concerning trends in trade lanes is crucial.

2.1 Classification of materials

Hence, the relevance of knowing how containers are moving is clear. Having an intuition of the reliability of the datasets improves the comprehension of how container flows are behaving and will behave in the future. However, before introducing a method that is able to capture these insights, a classification is necessary to identify different types of cargo.

A common way to classify vessels and their commodities is by scoping on the way the cargo is being transported. In general, cargo is either shipped by containers or as a bulk good. This classification will function as a basis, from which the quality of the different datasets will be determined later on in the report. In this section both types of cargo are elaborated.

First containerized cargo is further explained. A (shipping) container is a standardized reusable steel box, which is used to store or move materials and products in the global containerized intermodal freight transport system efficiently and securely. Intermodal transport is regarded as the shipment of cargo and the movement of people (in this context: containers) involving more than one mode of transportation during a single, seamless journey (Jones, 2000). The container has become an essential tool within the maritime transport network. Containers have had a prominent role within the recent transport history, as pointed out by Rijsenbrij (Rijsenbrij & Van Ham, 2012). Its development has had had a positive effect on shipping times, reliability and shipping costs. Its specific designed ships and handling equipment make the container fast, secure and inexpensive. Typical commodities that are shipped by container are industrially produced goods and intermediate products (Grossmann, 2007).

On the other hand there is non-containerized cargo, mostly bulk cargo. A formal definition of bulk material is given in (De Grace, 1968). They define bulk cargo as free-flowing material that is either loaded by shovel, pump, bucket or scoop. Moreover cargo is said to be stowed in bulk, when it is stowed loose instead of being first packed in containers. In this definition, containers are defined as conventional sea TUE containers. The bulk material is shipped in designated ships; bulk carriers. Cargoes that are historically shipped in bulk vessels are petroleum related products, grain, coal, iron ore, scrap iron, raw sugar phosphates and sulphur.

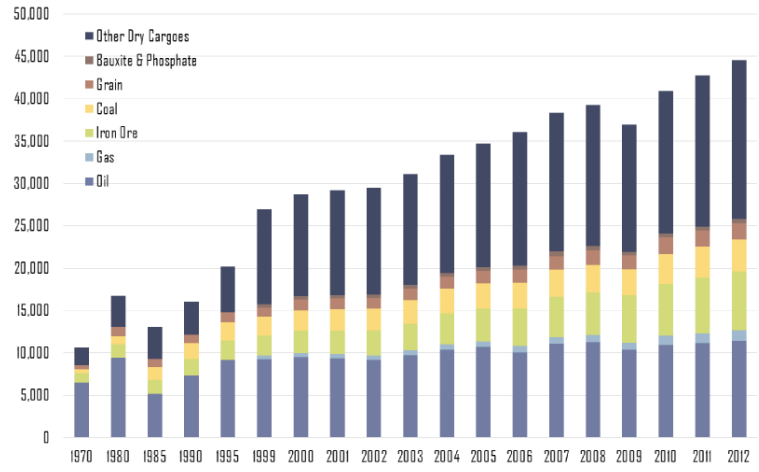


Figure 3: Ton-miles shipped of bulk commodities through the years globally. Source: UNCTAD Review of Maritime Transport (UNCTAD, 2013).

Figure 3 denotes the development of shipments of bulk material through the years. It shows that the amount of transported bulk material has increased over the last 40 years. Moreover, in the last ten years the amount of ton-miles of bulk material has increased with almost 50%.

This increase is caused by a strong demand of dry bulk. Research concerning the possible containerization of these materials is therefore urgent as the magnitude keeps on increasing.

Although historically dry bulk goods are shipped in bulk carriers, a shift in transport type can be observed. This is primarily caused by the trade imbalance between certain trading countries. A difference in the type of commodities imported versus the type of commodities exported leads to differences in the amount of containers entering and leaving a region. When this happens on a structural basis, this is called a trade imbalance. The U.S. for instance, exports a lot of dry bulk commodities to Asia transported in bulk carriers, while importing consumer goods from Asia transported in containers. This results in a large amount of empty containers situated in the Unites Stated. As a consequence, with an increasing supply and a constant demand, the freight rates drop. It is foreseen that also in the future trade imbalances between Asia and Western Countries will continue to exist (Diaz et al., 2011).

As a consequence of the drop in freight rates, bulk materials are being shipped in containers. Food importers are switching from dry bulk cargo ships to container vessels. In Figure 4 the two methods of transporting wheat are depicted. On the left wheat is put into designated bags, while the traditional loading of bulk vessels can be seen on the right. This tranformation allows transport companies to ship smaller quantities and at lower freight rates. As the wheat transport has been growing over the last years the transport method as been changing. Analysts estimate that up to 15% of the Australian wheat exports are now done through containers in 2014 (McFarlane & Saul, 2014). Brooks (Brooks, 2012) argues one of the biggest advantages of shipping in containers revolves around the allowed flexibility. While the minimum volume for a transport bulk carrier is tens of thousands of tons, containers allow tranporters to ship only 25 tons of wheat. Moreover, with the widening of the Panama Canal it is argued that containers can be dropped of at different locations. Hence, while the size of transported volumes might be constant, the containerized transport is expected to increase for these commodities. This is an example of a trend which is hard to observe using convential absolute trade volumes, but becomes apparent when analyzing container ratios. As a case, this is researched in Section 4.6.

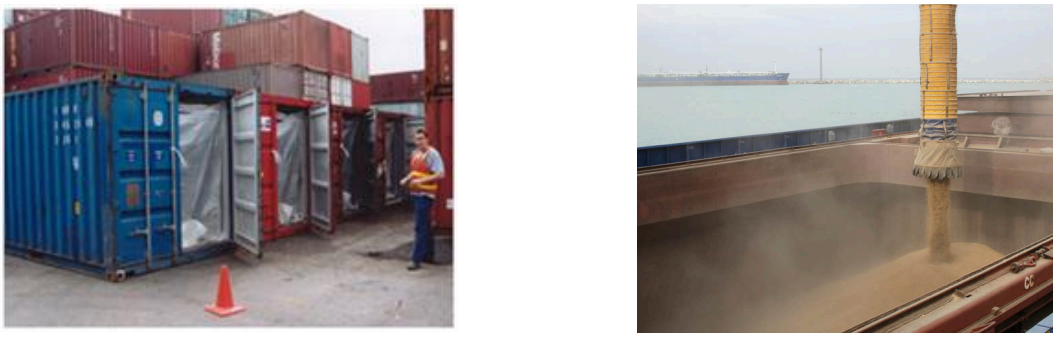


Figure 4: Transport of wheat by means of containers on the left (Nieuwsblad Transport NT, 2011) and in bulk vessels on the right (Maritime Sun, 2012)

The above classification can be summarized in a diagram, as is denoted in Figure 5. Cargo can be split into containerized transport, bulk transport and other types of transport (special transport like heavy machinery or offshore parts). Within the bulk group, one can distinguish two other classes: dry bulk and liquids. Moreover, it is expected that also on the containerized side, dry bulk is transported. This classification is on a very high level and purely based on transport types. As already mentioned in Chapter 1, this report will stay on a high level in analysing the behaviour of the trade lanes. However, as this is an exploring research into CR's, if the results demand a zooming in onto a certain commodity/region, especially in the dry bulk group, this will be done.

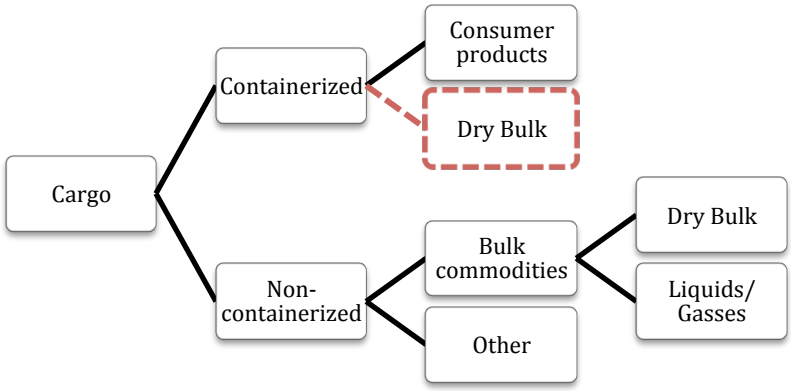


Figure 5: Classification of cargo based on the transport method.

2.2 Container Ratio

In the previous paragraphs a general framework is presented in which commodities can be placed, based on their transport method. This division is being somewhat obsolete for some materials, as explained. In this section a proxy is introduced which is able to assess the quality of datasets while considering the trends in containerization.

In this report the Container Ratio (CR) is being used to assess the quality of the datasets.

This CR is defined as:

$$CR_A = \frac{Total\ shipped\ volume\ of\ commodity\ \text{"A"}\ shipped\ containerized}{Total\ shipped\ volume\ of\ commodity\ \text{"A"}\ shipped\ containerized\ and\ noncontainerized}$$

Hence, the CR is a ratio that denotes the portion of a certain cargo shipped in containers. For example: shoes are expected only being shipped by containers. Hence, the CR of shoes is expected to be one. However, a commodity like rice is often shipped in bulk-bags which are sometimes loaded in a bulk vessel and sometimes in shipping containers. The CR of rice could be around 0.5. This implies that the assessment of the quality of data sources based on rice is harder without consulting expert judgements. Hence, the ratio summarizes the part of a selected trade flow that is shipped by means of containers. It is a ratio that can be calculated for a specific subset, a determined direction and tracked through time.

To our knowledge, no scientific study has been conducted specifically on the ratio of containers in determined trade lanes. However, there could be are a number of arguments for using the CR in the proposed context. First of all, the CR is able to visualize the relative trends in containerization as opposed to absolute numbers. Grossmann (Grossmann, 2007) states that individual groups of goods, like dry bulk goods and containerised foods will show different grow patterns when considering the trade volumes. This is confirmed in Figure 6. It denotes that the amount of dry bulk shipped worldwide has doubled in the last ten years to 600 millions DWT (Deadweight tonnage), while in the same time horizon the amount of shipped goods in containers has tripled to 200 million DWT. When only considering the absolute container volume number, this would lead to biased insights trend wise. The CR is able to isolate fluctuations in trade lanes and denote how likely containerization is for subgroups. In the previous section the containerization of dry bulk was discussed. It was deliberated that recently dry bulk commodities like wheat are expected to be transported more and more through containers. While the CR is able to denote these trends, trade volumes are only able to show absolute growth of the amount of containers within the trade lanes. Without considering the total volume of the trade lane for that cargo, the resulting discernment could be rather biased.

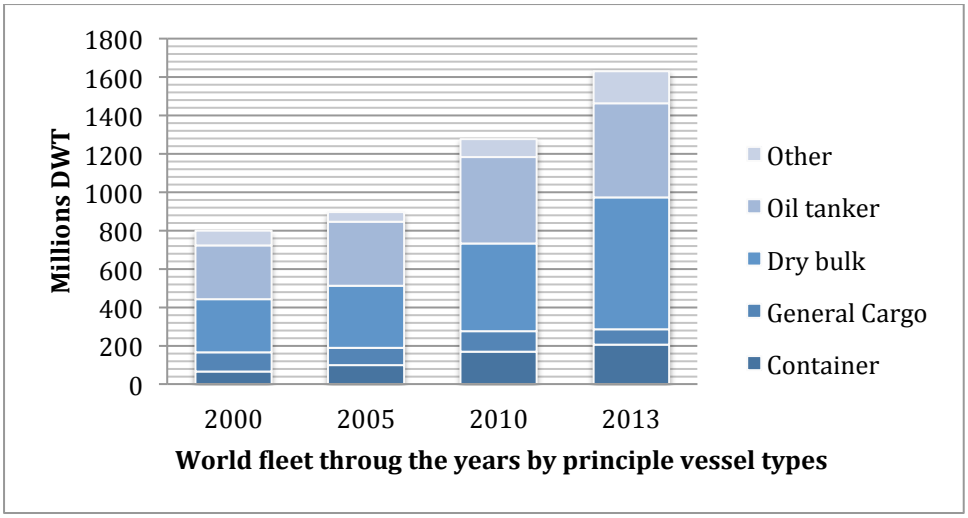


Figure 6: World fleet by principle vessel type through the years (UNCTAD 2013)

Another reason revolves around one of the goals formulated in this report. One of the main objectives is to assess the quality of the different sources of the dataset. As discussed in the previous section, there exists a group of commodities of which the type of transport (container or non-containerized) is known beforehand. Commodities like consumer fashion goods for instance, are solely transported in containers. By aggregating CR's on similar groups, one is able to assess the quality of the source. As we know the transport type, irregularities can be devoted to problems in the labelling or reporting of the source. Either way, it could point towards less reliable data sources. Exact methods on which reliability is assessed are discussed in Section 3.3.

3. Methodology

This chapter elaborates on the handling of the complexity of the data and the used methods for assessing the CR's of the datasets. There are number of steps that need to be taken before one can use the data for analyses. In Section 3.1 the problems associated with the size of the datasets are discussed and how these can be overcome. In Section 3.2 other specific procedures that should be performed on the raw customs datasets are discussed in order to use them for further analyses. The complexity of the data is reduced implementing the proposed methodology. In Section 3.3 statistical approaches are presented to assess the CR's of the datasets and determine their qualities.

3.1 Loading the datasets

For the statistical analyses the software package R (R.3.0.1, 2008) is used. The use of R for the statistical tests has several advantages over other languages, as also pointed out by Crawley (Crawley, 2013). R is a free open-source language, widely used by statistical experts around the world. It is used in every corner of the statistical academic world: for example medicine research, financial model validation, psychological testing, social experiments. R allows users to take advantage of the cutting edge applications it offers on a unrivalled number of topics. It relies on extensive documentation/examples and the number of users is still growing.

So in general, when conducting statistical research, R is a safe choice. Also in the presented scope, R is used to conduct the necessary analyses presented in Section 3.3. As already mentioned in Chapter 1, this report is only the first phase of a bigger project. After this report the ultimate model is expected to estimate CR's in the future. By using R in this phase of the process, statistical methods in later stages are less likely to be limited. Furthermore, for database management SQL is currently used. R allows users to set up a live connection to the database server easily, an absolute must to ensure reliable and easy accessible data mining.

Hence, the advantages of using R language to conduct the necessary analyses in this scope is clear. How to load the large datasets provided requires some more research. A conventional IT architecture for data projects R in combination with an external data source is given by Urbanek (Urbanek, 2013). He distinguishes three different sections in which data is being used in R. First of all there exist an (external) data source. It delivers the data using a connection through a database, for example SQL. This can be done both discrete and

continuous. The next critical part in handling the data is done by the data parser. It converts the original data format into objects that can be read by R. By doing so, the loaded data is put into a "data-frame". Now that the data in converted, it can be used for results (in whatever form that is). This is done in the processing phase. This pipeline is depicted in Figure 7.

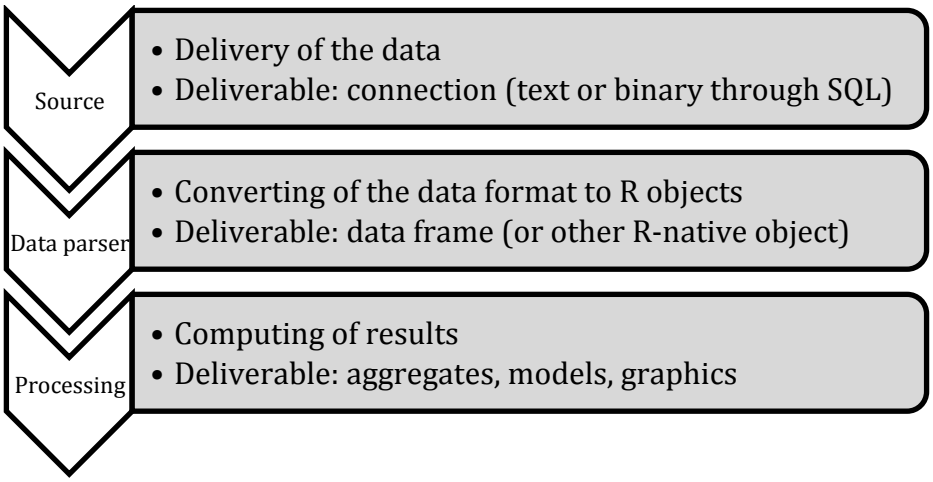


Figure 7: Data pipeline of a conventional R project (Urbanek 2013)

This is a classic and proven method of processing data through R. However, when handling big-data projects the proposed pipeline denotes some structural drawbacks.

The main drawback of the structure resolves around the way R uses its memory, as already pointed out by the founders of R (Ihaka, 1996). R stores all of the active datasets in its RAM, so in Figure 7 this implies that all the data from the data parser is stored in it active memory. In situations dealing with millions of observations at the same time, this design makes calculations slow and systems languid. Big data projects should therefore be handled in specific ways, circumventing these shortcomings. The most appealing approaches discussed in literature are listed below.

A rather straightforward method from a statistical point of view, is using sampling techniques or resampling techniques (Cormode & Duffield, 2010). Before loading the data R, one could select a random sample to process in a later stage. One could use sampling techniques to create models and/or validate the models created. This method is fast and popular. However for aggregation purposes, it is required to analyse all the data, especially when incorporating a weighting factor. Sampling methods are therefore more appropriate in the modelling and validation phase of the project. Therefor this method is not used for dataset loading.

A relative new methodology to overcome the shortcomings of R, or to deal with big-data in general, is by using parallel processing techniques. The parallel processing methodology

consists of at least three stages, as discussed by Urbanek (Urbanek, 2013). After the delivery of the data, the data is split. Every block of data is processed and calculations are being performed. After the computing phase, the calculations are combined to form the results. Figure 8 denotes the architecture of this methodology.

Numerous research has been conducted on the parallel processing of extremely large datasets. The most popular method of parallel processing is so-called Hadoop (Sudipto, 2010). It is a map reducing algorithm which can be used to perform chunk wise computing, even in a live environment. Although it is fairly popular and companies like Facebook and Yahoo are using it to process their data, its map reducing capacities can be questioned (Thusoo, 2013). This methodology has been researched continuously over the last years and it is able to handle the biggest datasets rather fast. It is able to load big datasets efficiently and has been used often in combination with R.

While the methodology seems promising, some drawbacks need to be pointed out as well. First of all, although research has been conducted on the accessibility of parallel processing, setting up the structure remains a complex job. Recently, dedicated software packages were released to overcome this problem. To cope with the poor map reducing techniques and to make Hadoop more accessible, warehouse solutions like Hive are introduced (Thusoo, 2013). However, the methodology still requires a fair amount of knowledge of the matter. More importantly, the proposed scheme is hard to incorporate in this report. As to be discussed in Section 3.2, the weighting of the current CR's is crucial in handling the data. This implies that for aggregations purposes, the whole dataset should be considered (source wise) and the splitting of the data should be done for every subset. Hence the first step of the parallel processing, the splitting, needs to be iterated for every subset. However, if the number of sources is bigger in the future or when not considering trade lanes but individual containers, the use of parallel processing would be inevitable.

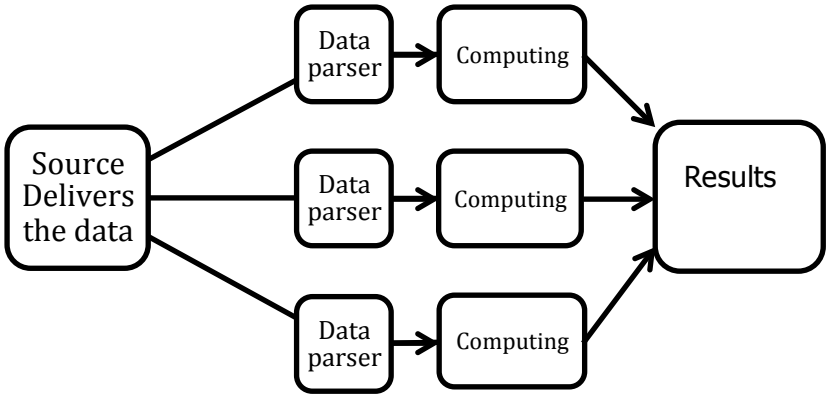


Figure 8: Parallel processing of the data (Urbanek 2013)

The last option is to apply some sort of aggregation at the dataset itself. A mapping scheme could be deployed to explicitly divide the dataset. Rosario argues that for applications where the files need to be considered as a whole, parallel processing techniques are not suitable (Rosario, 2010). He argues that there are general calculations for which parallel computing is appropriate, like word counting. However, other tasks need to consider the dataset as a whole, for instance weighting or aggregations. In these cases, the data structure could be aggregated rather than the data architecture. This method is ideal for datasets in the range ten gigabytes, whereas the explained parallel clustering techniques are more suitable for larger datasets.

A package that is able to implement this structure is "*ff"* package (Adler et al., 2007). This package contains a methodology in which the datasets are accessed on a file-base. Only the active parts of the dataset are loaded into the main memory. It keeps track of the metadata of the whole object, also the non-active parts. Hence, it only considers parts of the data at a time: only chunks of data are loaded upon request. The *ff* package could results in significant smaller RAM utilization, leading to a leaner model (GStat, 2010). The proposed method is in contrary to native R methods like the standard *read.table* function, which places all the files in the computers RAM. The *ff* package is able to handle a lot of different atomic types, including POSIXct, a common date structure.

It is argued that this method allows users to work with multiple large datasets while keeping the memory of the system clean. The lack of workable examples and the long unzipping times are considered as the main drawbacks. Moreover, in some cases the explicit

aggregation can be a problem, since it requires users to think about a possible separation of the data.

In this report it is chosen to use the latter option. The "*ff*" package is deployed to aggregate the data before working with it. The datasets are aggregated on their source. The R statistical software is linked to SQL query to perform the aggregation. Hence, the dataset is aggregated at the data source on the issuing country.

3.2 Dataset specific operations

In the previous section a scheme is presented to load the datasets into the R software package. However, the dataset is not yet workable and needs to be "cleaned". The most crucial handlings necessary to work with the data in a responsible manner are described in this section. The steps contribute in decreasing the data complexity.

One of the variables presented in the raw data is the Percentage of the Container Ratio. For every single entry, the percentage of containers involved in moving the specific good is listed. Hence, the majority of the CR's consists of zero's and one's. However, given that the entries have large deviations in weights, using this CR for further analysis would give a large bias towards higher CR's. Therefore, the observations need to be weighted based on their claimed weight in KG.

The arithmetic weight is used to correct for the bias flowing out of irregular entry loads. This is a popular measure to weight observations using a predetermined vector, as discussed in many books, like (Medhi, 1992).

The weight is defined as:

$$CR_{i,weighted} = \frac{\sum_{i=1}^{j} CR_{i,unweighted} \cdot W_i}{\sum_{i=1}^{j} W_i}$$

, in which CR_i denoted the container ratio of the original entry, i the selected entry and W_i the weight in KG of this entry. For feasibility reasons, the weights are normalized with the constraint $\sum W_i = 1$.

The biggest drawback of this method is its limited flexibility. Due to the changing sample groups defined in the denominator, for every different aggregation the weights need to be calculated again. Moreover, since the data consists of very specific trade data the aggregation can consist of millions of rows. Conventionally one would have to use a loop

structure to aggregate for specific lines and characteristics of trade lines. This is, especially in a big-data environment, a very slow, big coded and computational intensive process. A methodology that is able to overcome these drawbacks is the Split-Apply-Combine Strategy, as discussed by Wickham H. (Wickham, 2011). The proposed algorithm eliminates the extra code and decreases the calculation times for large datasets. The algorithm is basically a way of performing operations and computations parallel to each other by splitting the data in the first phase. It is important to denote the difference with the parallel splitting of the whole dataset discussed in Section 3.2. The Split-Apply-Combine splits the data after an aggregations is requested, not on a data-set level. The methodology can be used by installing the "*plyr"* package for R. (Wickham, 2011)

Another problem focuses on the labels of the trade lanes. Transactional raw custom's data, as mentioned by Versino (Versino, 2010), consists of a number of variables. Data fields may include a code classifying the commodity traded, quantity, value and date of the shipment, country of import/export and party names. For the analyses in this report the variables value and party names are not taken into scope.

The classification of the commodity is key in explaining and analysing the CR, as discussed in Chapter 2. However, the ways commodities are mapped are not consistent through the different datasets. Eurostat uses its own NSTR mapping to group commodities. The UN dataset, out of scope in this report, uses its own SITC mapping for trade flows. However, the majority of the datasets are classified using the HS system (World Customs Organization, 2012). They state that HS coding has become the standard taxonomy for commodities in a majority of issuing countries for trade associations, statistical offices and customs.

The used mapping in this report is different. Within Seabury, commodities are classed into G-codes. The reason for this difference in mapping originates in other advisory roles Seabury performs in the air transport sector. As a results, the classification of dry bulk materials is granular. The G-mapping has the following structure. Every trade lane has its own G4 code, which denotes the type of commodity shipped. This G4 Code is based on industry-standard HS8-10 coding of the container or bulk shipment. A G4 code falls in a G3 group, which falls in a G2 category which eventually falls in a G1 group. Hence, there are four levels in total in the G coding. In the highest G1 level based on industries, there are 12 classes. On the next level (G2) there exist 60 categories. Going one level deeper, there are 339 G3 classes. Likewise, there are almost 2000 G4 codes to classify goods.

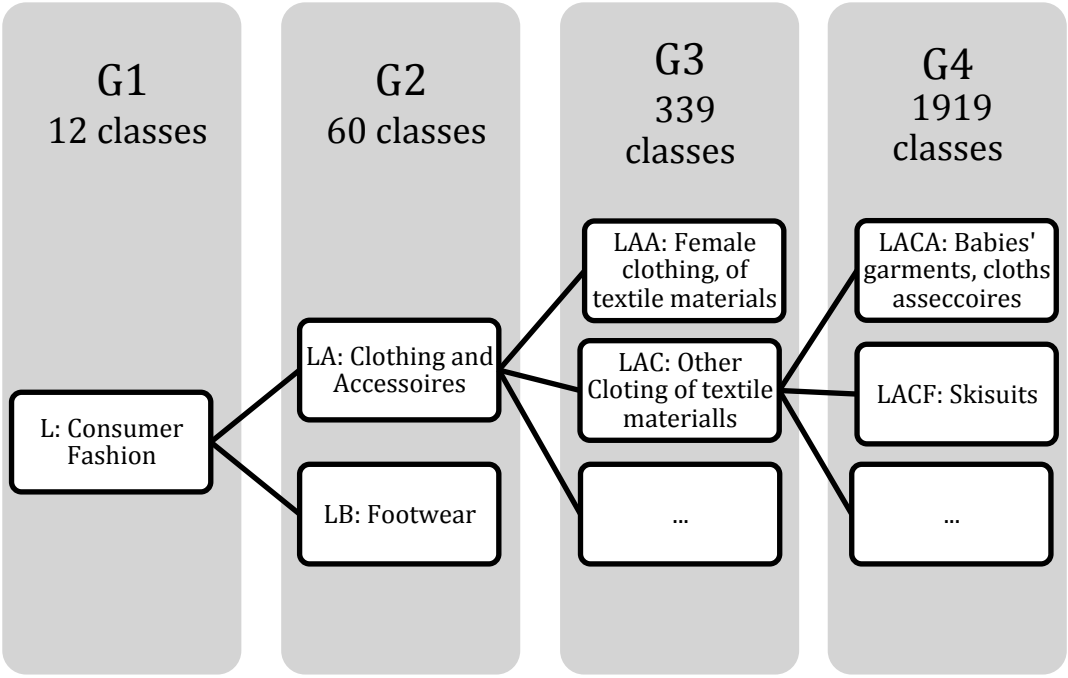


Figure 9: Example of classification using G coding.

An example to illustrate the G-mapping is given. One of the G1 codes is the "L" group. This group denotes the "Consumer Fashion" commodities. One of the G2-codes belonging to "L" is the "LA" group, representing "Clothing and Accessories". Going even less granular, one could find the G3 group "LAC", representing "Other clothing of textile materials". On the most specific G4 code, one could find the "LACF", which denotes "Skisuits". This classification example is depicted in Figure 9. Hence, when aggregating on the LACF code, one is able to see how the skisuit trade is going and what the CR of this product is. In this report, for feasibility purposes analyses are only conducted to a G2 level. Only if the CR of a G2 code is behaving different than expected, deeper levels are analysed to find out if this is due to a deeper classification.

This implies that the HS codes should be translated in corresponding G-codes. The standard datasets with HS6 (or even HS6-10) codes is translated in fairly specific G4 codes. As one of the goals is to research the behavior and trends of CR's on a more global level, the codes are re-aggregated in a later stage to G1 classes. The Eurostat dataset (Eurostat, 2014), discussed in Chapter 4, has its own NSTR classification. Their classes are also translated into G mapping, but for transparency reasons the underlying NSTR groups are displayed in several analyses.

The next phase if the data handling, focuses on the practical cleaning up of the data. There are number of operations which make the set usable for quality methodologies discussed in Section 3.2.

The first operation resolves is the splitting of trade containers and traffic containers. Traffic is defined as containers in which the noted destination is not the final destination, hence transhipment is occurring. The way these transhipped containers are recorded are not the same for every dataset (Fleming, 1997). However for the analyses conducted on CR only transport is of interest. Traffic containers are therefore mutated.

In addition, also empty containers are removed from the datasets. Containers with a reported weight/value of zero are dismissed as they can bias the amount of containers shipped. For CR calculations, they do not have an effect since this is calculating using the relative weight of the selected subset.

The final phase of the cleaning the data are small operations. The variable names are made consistent throughout the different datasets. A handful of observations with unrealistic high values are removed. These handlings improve the reliability of the dataset and make the codes modular between the different data sources.

3.3 Assessing the quality of the datasets

In the previous sections it was explained how the dataset was handled to make it suitable for statistical testing. First the loading scheme was explained in Section 3.1 and in Section 3.2 the operations necessary to make the dataset usable were explained. In this section, a methodology is introduced which is able to analyse how CR's are moving. Hence, given the cleaned up data set, how are the CR's investigated.

First of all, the historical relevance of the different CR's are plotted. The imported and exported trade lanes are separated to isolate trade imbalance differences mentioned in Section 2.2. On a data source level, the imported and exported CR's are denoted. It is researched whether the CR's move in trends through time or if the CR's move randomly through time. It is expected that time has a significant effect on the CR. Not only can there be economic incentives to move cargo in bulk or in containers among transport companies, also seasonal effects of certain commodities in subsets could have an impact on trends of CR's.

Before comparing the datasets with each other, it is necessary to get a thorough understanding of how container ratios behave and what differences exist among classifications. It is important to get an intuition in the effect of different variables (like the cargo shipped and the direction of the transport) has on a CR. Therefore, existing

assumptions are tested on the United States dataset. Although this does not necessarily lead to conclusions for the other datasets, it is assumed that the general findings are in the same line. The main advantage of the US is its size; given its large import and export flows the CR's are "reliable" (not biased by certain shipments) even on a very specific G4 level (Iseman, 2014). Appropriate summary statistics are plotted and the results are discussed and interpreted from a transport point of view.

After assessing the influence of commodity type and direction on the CR, the other datasets are considered. It is researched how the influence of the mentioned variables is between the different sources. Furthermore, the validity of results are interpreted. For instance CR's of consumer products are expected to equal one. Deviations are researched and, if possible, the underlying reasons for deficits are subjected.

After that, another method for reliability of the different sources is performed. In a trade lane, volume is recorded in two places; the exporting country under exports and the importing country under imports. Under normal circumstances these two volumes should denote the same numbers; hence the CR's series should lay on top of each other. If a deficit exists, at least one of the sources is reporting in a different way. It is impossible to validate the series since the "real" trade numbers are not given out, but by comparing the sources against multiple other sources one can create an intuition regarding the reliability. Again, the reliability is determined based on consistency. Given the aggregation in the Eurostat data, it is not feasible to perform a similar analysis on this dataset.

Finally, the ability to spot underlying trends using the CR is tested. The intuition described in Chapter 2 is tested: the existence of a correlation between the shipped volume and the CR. The United States historically exports a lot of bulk material and imports goods that are shipped by containers. This mismatch is called the "Trade Imbalance" and results in a large number of empty containers on US soil (Robinson, 2007). There are several ways to get rid of this imbalance. One way is to sell the containers locally, for transport purposes. However, it is obvious that this is not a very sustainable solution as this market gets saturated. A more straightforward solution is using containers to ship bulk material. Although this might been not economic feasible in the past, the containers available at a rather high discount. It is expected that transport companies supply container shipments for low freight rates to ship bulk material out of the country (Gurning, 2007). Hence, the hypothesis is formed that on average, the Container Ratio decreases if the weight of the shipment increases for certain groups. This expectation is tested using the different data sources. For more background information regarding this subject, consider Chapter 2.

The above statement is tested using the provided datasets. Due to the large number of observations, fitting a line on the data points using ordinary least squares techniques is not

feasible. This research focuses on classic bulk materials that are shipped by bulk carriers. Since considering all materials is not feasible, two dry bulk groups are selected: the "GA" group representing the "Foodstuffs & Beverages for human consumption" and the "FF" group representing "Other chemicals".

A first step to get an idea of the possible correlation is to order the weights of the trade flows and calculate the average CR. However, this is not a fair comparison since it does not take the type of commodities into account. The smaller trade volumes are shipped by container given the format of the commodities. However, this relationship does not give any information concerning the hypothesis mentioned earlier, because trade categories are not aggregated. Therefore, it is not a fair comparison on an entry level since it does not show any relationship between smaller mass flows and higher containerization for comparable commodities.

However, the relationship between the weighted CR and the summed monthly weight shipments on a group and import/export trade lanes can be denoted more exactly. The expectation is that for months where less is shipped, the CR is higher. To test whether there does exist a significant relationship between the weighted CR and the total shipped volume, the standard linear equation is regressed on the observations. The resulting model is $\hat{y} = Xb$, in which $\hat{y}$ denotes an estimate of the CR and x the explanatory total weight of the shipments. The b denotes the estimated coefficient (or estimator), and hence yields the performance of the regression. The significance of the estimator should be considered closely, since it gives information if the effect is neglect able or not. Not that there still may be a substantial amount of variation around the regressed line. This is an indicator that other variables have an effect on the CR too, which are independent from the volume regressor. The regressions are checked upon heteroskedasticty and endogeniety and the mentioned variation of the denoted is checked using the R-squared.

4. Results

In this chapter the results are denoted of the analyses conducted on CR's of raw customs data. In Section 4.1 the used data is shortly introduced. In Section 4.2 some results regarding the handling of the data are denoted. Section 4.3 denotes some remarkable results when the CR is researched coming the US data source. The influence of categories on a high level is researched, the CR of a fully containerized category is analysed and the influence of transport destination on CR looked into. In Section 4.4, the CR trends for the different data sources are compared to each other. A cross validation between the sources is denoted in Section 4.5. It is researched whether exporting trade flows are equal to their importing side when considering the container ratio. The chapter concludes with Section 4.6, in which the relationship between trade volumes and CR are researched for selected subcategories.

4.1 Introduction of the data

In this section the provided data is described. For a methodology to handle the raw customs data, refer to Section 3.2 and Section 3.3.

The analyses are conducted using different datasets. The datasets originate from different national customs. Table 1 denotes the different datasets and some of its characteristics. The datasets have a large number of observations or rows. Each row represents an entry by customs. Basically, these entries are monthly data of the different categorized transport flows from one place to another.

Dataset name	Description	Number of trade lanes	Number of variables	Timeframe
US_C_Ratio	Customs data of the USA	16.059.472	17	11/2008 - 06/2014
JP_C_Ratio	Customs data of Japan	6.201.559	13	01/2008 – 06/2014
Eurostat_CR	Collected data of EU	31.597.856	12	01/2000 – 06/2014
ES_C_Ratio	Customs data of Spain	118.161	12	01/2008 – 11/2013
UK_C_Ratio	Customs data of the UK	200.768	13	01/2005 – 06/2014
TW_C_Ratio	Customs data of Taiwan	205.381	13	01/2008 – 06/2014

Table 1: An overview of the considered data

The data is aggregated on their source, as explained in Section 3.1. A method to extract the data from the different sources, is introduced by Versino et. al. (Versino et al., 2010). The Eurostat data denotes trade information from a collection of EU countries (Eurostat, 2014). Other customs data considered in this report are the US, Japan, Spain, the United Kingdom and Taiwan. To our knowledge these countries are unique in their containerization information provided. The trade data is divided in cargo transported through containers and cargo transported outside containers.

This allows us to calculate a CR, in contrary to the gross of the customs data worldwide. The data was retrieved from their sources in July 2014. Note that the time horizons of the data sources are not consistent, this makes conclusions based on direct comparisons more complex.

4.2 Handling of the data

Given the introduced in the previous section, some of the results revolving around the handling of the data are presented in this section. The explained methodology of Section 3.1 and 3.2 are deployed on the dataset in order to clean up the raw data sources. The global findings and some helicopter view statistics are presented. This allows to get a better understanding of the CR and how it is influenced.

The loading of the datasets is deployed as explained in Section 3.1. To illustrate the performance of the "*ff*" package a simple aggregation is performed, calculating the total export and import container ratios. Recall that the *"ff"* package is able to effectively aggregate the dataset into the different sources in R, making statistical calculations conceivable. As a result, the calculation times are feasible, denoted in Table 2. This implies that the proposed methodology leads to workable waiting times and makes further research on the datasets possible. Denote the non-linearity in these calculations: bigger datasets require longer going through than smaller datasets. This is caused by the way R reads and writes in vectors; bigger vectors require more computational power when processing, than smaller ones.

Dataset	Amount of observations	Calculation time
US_C_Ratio	16.059.472	614 seconds
TW_C_Ratio	200.768	22 seconds

Table 2: Illustration of the proposed data loading scheme, when applying a simple aggregation.

In Chapter 2 the weighting of the original CR's is discussed. It is expected that weighting the original container ratio had a downward effect on the CR. This is confirmed in Figure 10. The green line denotes the original unweighted CR's and the red line denote the weighted observations. The trade lanes with heavy and big volume raw materials are now weighted, and have a bigger impact on the CR of the subgroup. Since these raw materials are shipped by bulk vessel, this leads to a drop of the CR of the subset: the unweighted CR was around 80% but is adjusted to 20%.

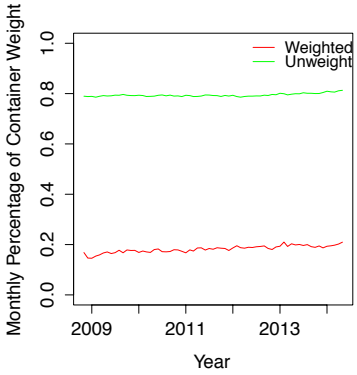


Figure 10: The influence of factor weighing on CR's when aggregating.

When zooming in on the corrected CR, it is necessary to aggregate on "Import" and "Export". Given that the United States is exporting a big amount of raw agriculture commodities, the CR of the exporting side is traditionally lower than the CR of the importing side. However, Figure 11 denotes that over the last 5 years there as been a change in leading CR. Trends can be observed, hence some sort of historical context influences the CR. The importing CR has some visual seasonality and a downward trend. Halfway the year 2014 it has reached a level of 0.18. The exporting side seems to be less influenced by seasonal effects and has an upward trend. Halfway 2014 it has reached a level of 0.22.

This change of the leading series could be influenced by a number of causes and will be explored in the remainder of this chapter. A note that could be made quite often in the report, is that the presented CR figures are not able to visualize the underlying mass flows. Hence, a declining CR could be caused by fewer shipments in containers but also by a drop in a certain specific travel flow. Hence, the illustrated figures should always be interpreted with care; they only display relative movements.

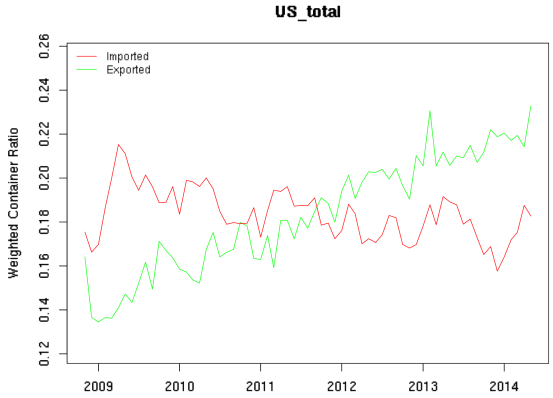


Figure 11: The CR of the import and export United States customs

4.3 Analyses conducted on the US custom's data

As discussed in Section 3.3, first summary statistics are calculated for the US data source. An intuition will be created, from which the other data sources can be evaluated. The effect of categories on the container ratios on both a G1 level as a G2 level and the regional effects are researched. The lucidity of the G mapping is discussed in Section 3.2.

In Section 4.3.1, deeper aggregation levels will be used, based on the categorical mapping of the trade flows. This will allow us to investigate the effect of the type of commodity shipped on the CR in the Unites States. When compared with prior knowledge of the field, a first impression of the reliability of the results (and the underlying dataset) can be formed. Firstly, a granular aggregation based on the G1 classification is made. After that, a selection of G1 groups is further researched on a G2 level. The correlation between the classes is calculated and findings are interpreted. In Section 5.2.2 one of the findings in the Consumer Fashion Good category is explored on the deepest G4 level.

4.3.1 Categorical variables

In this section the CR of the G1 group of the US is further researched. As already mentioned, the G1 grouping is the most granular grouping within the G classification. Containers and shipped goods are classified based on their global industry. Hence, it is expected that the

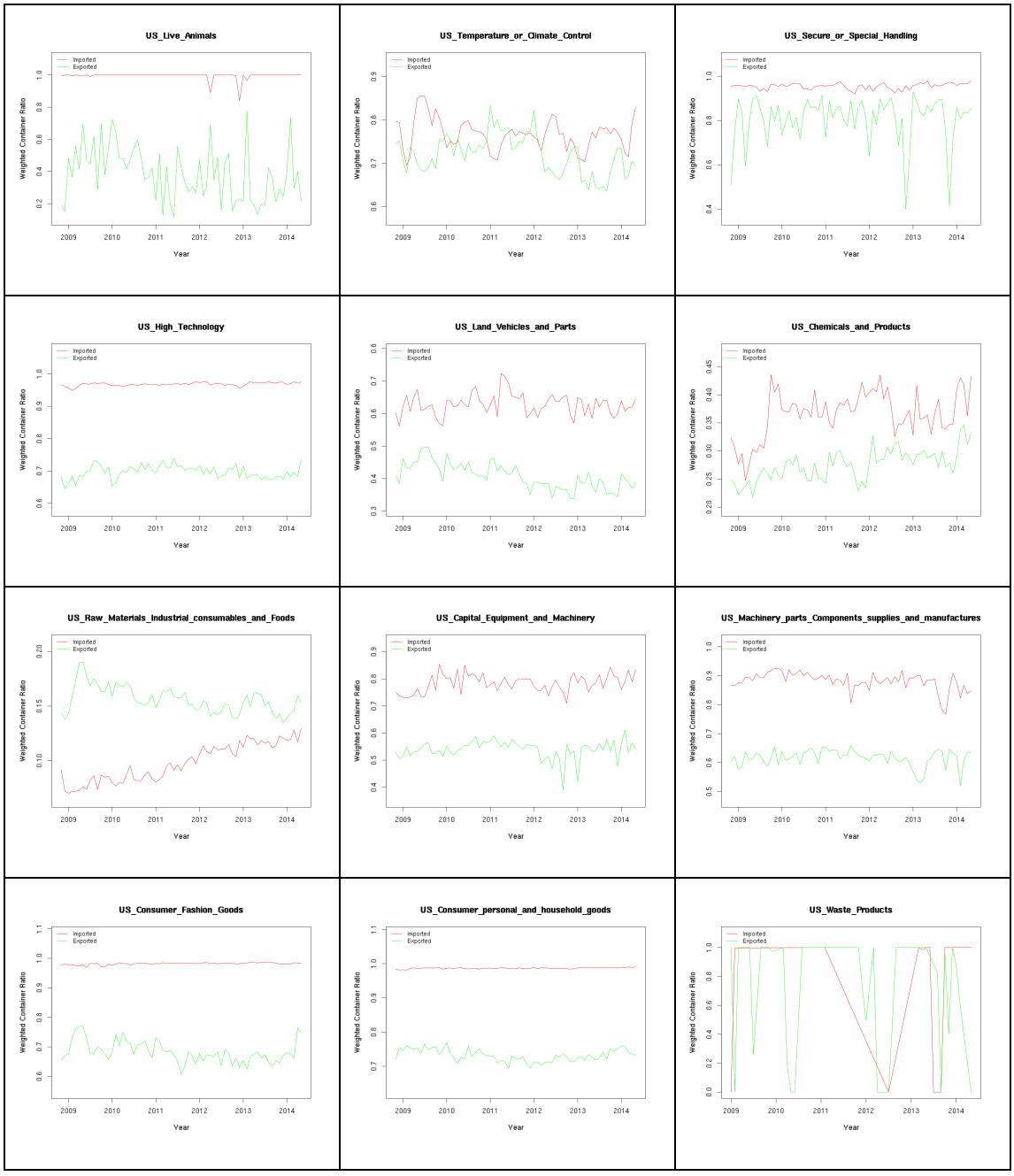


Figure 12: The CR's of the import and export stream of the different commodity classes for the United States dataset. The figures represent the different G1 groups, namely (left to right, row per row): A, B, C, D, E, F, G, H, I, L, M and Z.

levels of the CR's vary to a large extend, given that some commodities are shipped by container and some in a bulk vessel. For a list of the letters of the G1 classification and their group names, consider Appendix A. The weighting scheme of Chapter 3 is deployed.

Although a first aggregation is done based on their G1 code, also the imported and exported stream should be separated. Figure 11 could be regarded as a first indication that the importing side of a commodity could have a different CR compared to its exporting side. This

is caused by a number of factors. First of all, on a G1-level the grouping of products is still rather granular. It is very well possible that the type of products (and therefor also their CR) is different on importing side and the exporting side. Counties could have a difference on the ratio of importing and exporting trade lanes. The United States for instance has a large trade imbalance; on the transpacific trade lanes the ratio of container was 1:2.6 for respectively exporting versus importing in 2006 (Robinson, 2007). This, combined with the difference in trade volumes leads to differences in the importing and exporting CR's within the same groups. So in the end the data set is aggregated on their date, G1 code and direction using the "*plyr"* package.

The results of the analyses can be observed in Figure 12. Although all of the figures contain both the importing and exporting flow of the specific G1 group, a comparison between the groups should be conducted with care; the values on the y-axis representing the CR is different among the figures.

A first general remark that could be made concerns the large fluctuations of some of the CR's. A relative high volatility could be seen in the groups B, C (export) and Z. For the groups C and Z this is caused by the small amount of shipments/volume in this specific class. When very few shipments take place in a specific subset, its CR influences the total CR to a large extend. It could be argued that this has a negative effect on the reliability of the CR.

Furthermore, seasonal behaviour seems to presents in a number of subsets. Within the groups B,E and H the time of the year has a visual significant effect on its CR. In the B group, representing temperature controlled containers, this seasonality was expected given the fruits transported (Jedermann, 2013). However, very few statistical tests are able to formally show this seasonality, especially if the amount of observations is limited.

From some categories it is expected that the CR is relatively high. Groups with the G1 code A (the live animals), L (Consumer Fashion Goods) and M (Consumer products) should have CR's of close to 1. However, especially at the importing side, the levels are lower than expected. For the "Live animals" group the imported CR is around 1 but the exported CR is significantly lower at around 0.4. Also for the consumer products, groups L and M the imported CR is around 1 but the exported CR is around 0.7. In general, this could be caused by a number of factors. First of all the G1 group could contain product that pollutes the rest of the CR. A specific product with a relative a high weight that is shipped a few times but with a large weight could lead to a bias for the rest of the groups. Furthermore, outliers and small trade lanes could have the same impact. For the "Consumer Fashion Goods" group, the low CR is further analysed in Section 4.3.2.

The amount of upward and downward trends throughout the classes is rather limited. Neglecting the white noise in the series, only four series do not show stationary behaviour. In

the G group the imported trade flows have an upward trend; in 2009 the CR laid around 0.05 but recently it has inclined to 0.13. On the other hand, the exported trade flows have a downward trend from 0.19 to around 0.14. This could be interesting, given the large mass flows involved in this trade lane. A further analysis of this trend is conducted in Section 5.2.2. Again, the series do not display the trade volumes. Although the CR of the imported side is lower compared to the exporting side, the absolute number of containers is higher on the exporting side. The other series do not have a trend on a visual basis.

When interpreting these results, this could have several reasons. First of all, the majority of the commodities are not shipped differently over the last ten years. Especially for categories where the CR's are high the CR has not changed dramatically. This makes sense, since industrial made products originally put in containers will not be shipped differently. Secondly, some of the CR's are known to be labelled incorrectly. Within Seabury it is known that the import ratios of the US, especially highly containerized cargo, cannot be trusted. One of reasons lies within the labelling. Whenever imported containers are directly put on trains, including trailers, they are not considered as containerized. Also some states have inconsistent ways of separating import and export. These historical phonemes are still not corrected.

The same analyses as above are conducted, only on a more detailed G2 level. Given that there are 61 G2 groups, only a selection will be displayed in the report. The same aggregation as in the G1 group is used, again with weighted CR's. By going on a deeper level, the relationship among the different groups can be observed together with the CR's compared to their higher level CR series. The section consists of two parts. First two groups are selected for a deeper visual analysis. After that, the correlations between the G1 groups and G2 groups are explored. This will give an idea of the consistency of CR's of a group towards its subgroups, a relationship that is often assumed to be present.

Only one G1 group is discussed here: The exporting trend in the "G" group, representing the "Raw Materials, Industrial consumables & Foods". The G group is selected because of its large weight in absolute mass flows. Hence, the influence of its CR is supposedly significant.

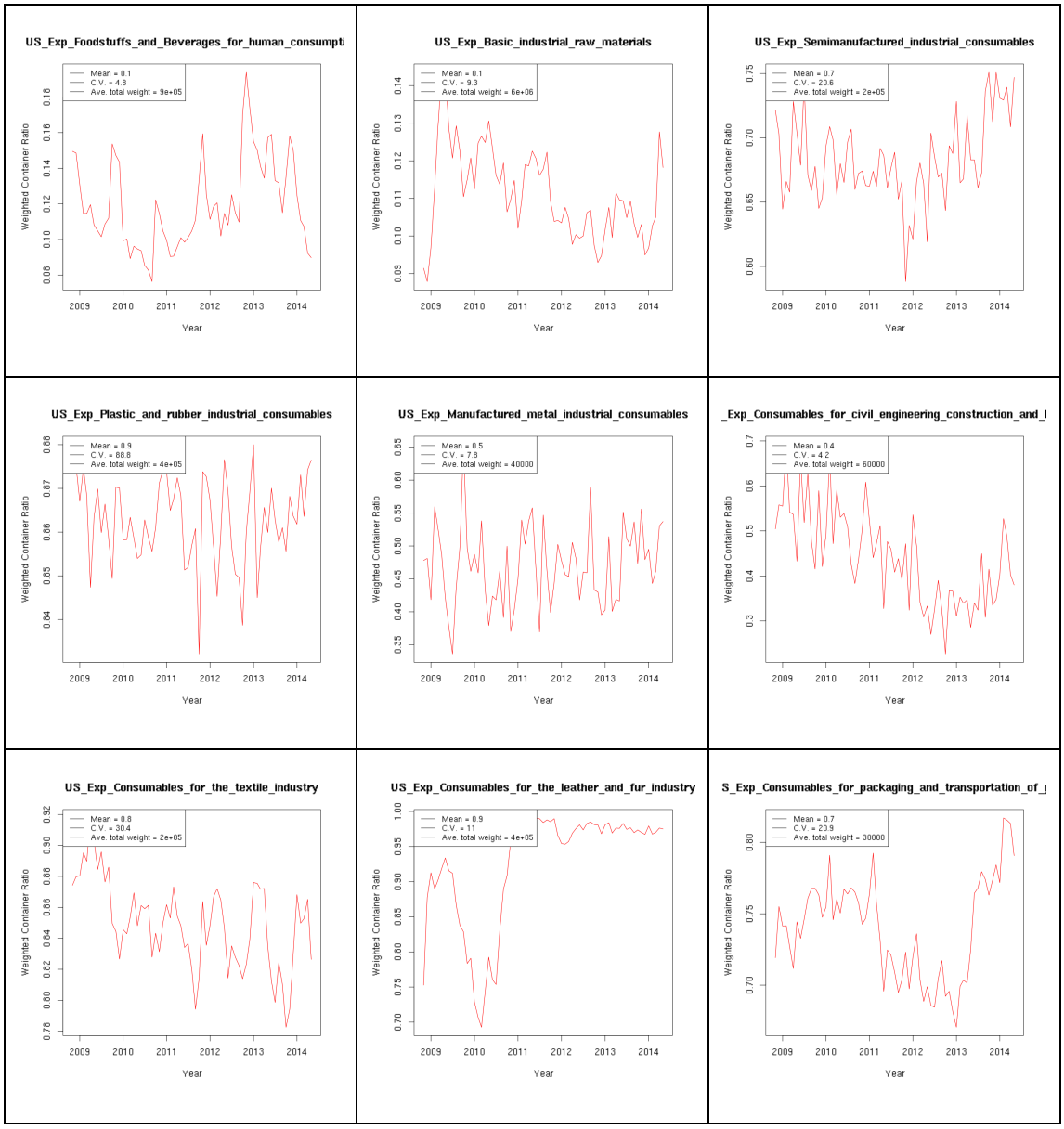


Figure 13: The exporting CR's of the G2 groups in the Raw Materials, Industrials consumables and Foods class "G" in the United States.

Hence, first the CR's of the 9 G2 codes in the "G" group. In Figure 11 it was denoted the CR of the exporting trade lane in US was slowly decreasing thru time. When observing the CR's of individual G2 classes, denoted in Figure 5, some general remarks can be made. Again, the CR's are highly volatile. Moreover, the average shipped weight of the classes seems to fluctuate throughout the different classes. In the GB, representing the Basic Raw Materials, group for instance entries had an average weighted volume of 5000 tons, where as other groups have 30 tons shipped on average. Presumable these differences have also its effect on the global CR of the G1 group. A group with a relative high weight will have a large effect on the group CR compared to group with a small weight. This is also tested in the last part of this section.

A closer look at the figures reveals differences of the level of the G2 groups. Only GA, representing the Foodstuffs and Beverages and GB have a rather low CR. Other groups in this category have a significant higher CR. This indicates that in the other groups the majority of the products are shipped by container. However, the effect of GB's mass removes this insight on a G1 level.

When considering the trends, only the GB and GH seem to have the same direction as the G1 trend discussed in the previous section. The other groups seem to have a volatile stationary behaviour. This is remarkable, given the fact that it is expected that similar products should have similarities in their CR as well. Apparently, within a category CR's do not necessarily have a high correlation.

In the previous paragraphs it was stated that a small group of G2 groups seemed to lead the G1 group to a large extend. This statement is formally tested by computing the correlations between the CR of the G1-classification and the CR's of the G2 classes. The results will give a rough indication concerning the explanatory capacities of the CR of the G1 group for the underlying G2 groups. The relationship is tested in the G-group (Raw materials, Industrial consumables and foods), F-group (Chemicals) and the K-group (Machinery parts. Components, supplies and manufactures).

		G2-Code									
		A	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**J**
G1-code	**G export**	-0.364	0.932	-0.074	-0.085	-0.035	-	0.453	0.463	-0.393	0.079
	G import	-0.239	0.88	-0.635	0.781	-0.22	-	0.26	0.191	0.209	0.65
	F export	-0.167	-0.13	0.164	0.018	-0.212	0.99	-	-	-	-
	F import	-0.176	-0.13	0.164	0.382	-0.189	0.996	-	-	-	-
	K export	0.323	0.19	0.036	0.044	0.092	0.492	0.345	-0.018	0.351	0.655
	K import	-0.329	-0.13	-0.079	0.178	0.045	-0.343	0.154	-0.191	0.264	0.815

Table 3: Table of correlations between G1 and its G2 groups of the US dataset for the "G", "F" and "K" groups.

The results of the correlation analysis are denoted in Table 3. The rows represent the G1-groups, aggregated for import/export and the columns represent the G2-codings. The correlation between the G1-group "G" and the G2-group "GA" for example is -0.364. Unknown correlations and not existing groups are denoted by a "-".

When observing the results, some interesting characteristics come to the light. There seems to be a big amount fluctuations within CR's of the data. When looking at the G-group, on the export side, the highest correlation of 0.932 can be found with the "GB" group. This is also denoted on the import side with a correlation of 0.88. The Basic Raw Industrials have the highest weight of the G2 groups, especially on the exporting side, denoted in Figure 5. While on the export side other G2-groups have little impact, the "GD" and "GJ" groups also have correlated CR series on the import side. Again, these are groups with a fairly high trade volume.

Similar behaviour can be seen in the F group Chemicals. The absolute leading G2-group is the FF-group (Other chemicals); on the export side the correlation is .99 and on the importing side 0.996. Again these are the groups that have similar ratios considering the shipped weights. Other G2 groups do not have correlations that have same order of magnitude.

Finally, when considering the K group, no G2 classes have correlations that "peak out". For both the importing and the exporting side, the "KJ" group have the highest correlation with respectively 0.655 and .815. This is again in line with the higher weights of these subgroups, denoted in Figure 6. The lack of consistency among the CR's is striking within this G1 group. Almost none of the G2 classes have high correlations with the G1 group.

Hence, based on the findings of the correlation table, it can be argued that there exists little correlation of the CR between the G1-group and G2-groups in general. However, for the classes with a very high weight, the correlation is extremely high. The CR of the G1 group is therefor largely influenced by those G2 classes that have large mass trade lanes. Although correlations give a rough indication of existing relationships, they work best in stationary series. Also the (relative) small number of observations could lead to inaccurate statistics.

4.3.2 Fashion goods

In Section 5.2.1 some of the CR's were lower than expected. The G1 group "L", representing the "Consumer Fashion Goods" turned out to have a CR of around 0.7 on the exporting side. This is confirmed in the isolated series of Figure 14. This is significantly lower than a value of 1, expected for fashion products since they are shipped in containers (Chen, 2012). In this section, this result is further analysed. Hence, a lot of other categories could be selected, but the general methodology and findings will give an indication how such low CR's are caused.

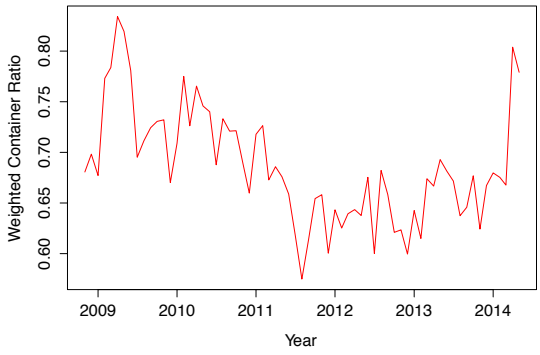


Figure 14: CR of exported trade flow of the "L" group, representing the "Consumer Fashion Goods".

There are several possibilities to research where the deficit is coming from. The CR's contain a mix of 0's (bulk shipments) and 1's (containerized shipments). By only considering those entries with that have a fully non-containerized trade flow (the "0's"), the origin of low level will come apparent.

First, the destination of the non-containerized transport flows is investigated deeper. The results are denoted in Figure 15. On the vertical axis the total shipped volume in KG is denoted. The horizontal bars indicate the countries of destination in the form of a code. The exporting locations are mainly within Central and South America. Guatamala (GT), El Salvador (SV), Honduras (HN), Chili (CL) are four of the five most intensive countries. Only Great Britain (GB) is.

The reason for this results is most probably mislabelling of the commodity. The fact that the vast majority of the destinations are in the same continent could indicate that the labelling of the trade flows is done incorrectly. If a trailer, truck or rail, is shipped it will automatically be labelled as non-containerized. The commodity however, could be shipped in a container. Hence, while the product is being shipped in a container, it is being reported as non-containerized. Experts argue that this is a typical US characteristic. This behaviour should not be denoted in the other data sources.

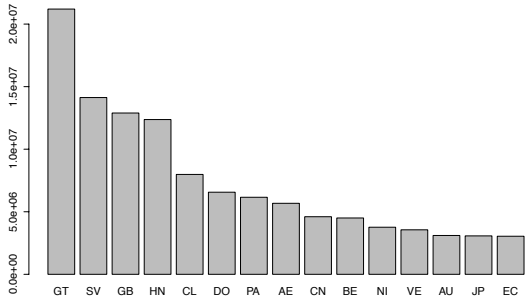


Figure 15: Countries containing the destination of the non-containerized "L" trade volumes.

Besides considering the destination of the commodities, also the shipped material could be analysed. The results of this analysis can be observed in Figure 9. Hence, it shows that the majority of the non-containerized CR's are caused by the LAEH category. This category is described as "Articles of apparel and clothing accessories, of plastics". Though the G1 category is highly containerized, it is possible that these materials are shipped as a bulk good. Especially the plastic is supplies could be shipped as a bulk material. However, the volume is quiet low for bulk shipments. More likely is the wrong labelling as mentioned earlier.

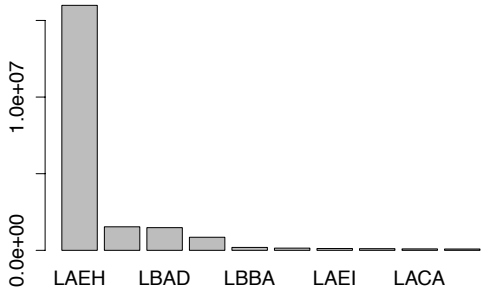


Figure 16: G4-categories that contain highest amount of non-containerized trade volumes

In the end, the lower CR is caused by a subset, which has other characteristics than the rest of the group. Especially the handling and transport of the plastics could be done in bulk vessels or in the form a trailer on a vessel. Hence in general, the low numbers do not indicate a lower reliability but it creates a noise in that category. It is possible that a very specific G4 trade flow with a substantial weight, leads to a bias throughout the whole category.

4.3.3 Regional Analyses

In this section the CR's will be analysed in the light of destinations and origins. In the previous section, analyses are conducted based on the categories of the transport lanes. The regional character however, is assumed to have some explanatory capacities as well.

The seasonality observed earlier, can also be denoted when looking at regional aggregations. Strong seasonality for instance can be observed when looking at the CR's of Central America. This is closely related to the commodities traded in that specific region, mainly temperature related products.

There seems to be little grouping possible within this variable, based on the CR. The trends and levels of the CR's are different for every region. It is hard to spot consistency on this level of analyses. Regions that are geographically close to one another, show completely different CR's. It is likely that there exists a high correlation between the type of commodity and its destination, an observation substantiated by literature (Kaluza, 2010). Kaluza claims that grouping ports into regional clusters is rather complex. Hence, the effect of the regions is rather biased also when zooming out. This observation leads to the assumption that the added explanatory capabilities of the regional variables on a group level are small; one would have to explore specific G4 codes (or even deeper) to find regional differences.

4.4 Cross country customs data comparison

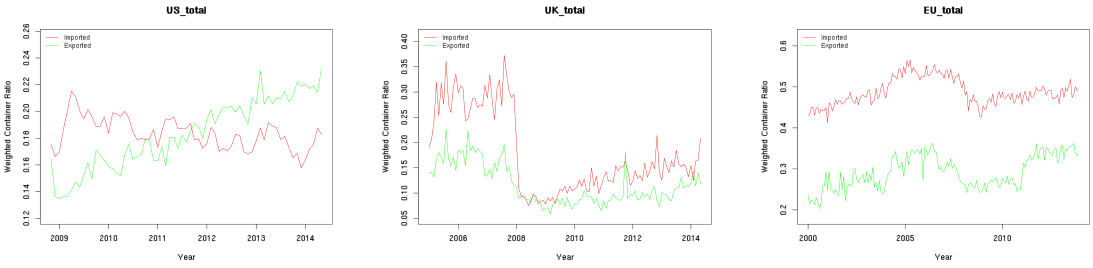
In the previous section an in-depth analyses was conducted on the US data source. Some of its explanatory capabilities were researched and the different trends were interpreted. The next phase is to see how the datasets relate to one another. In this section the CR's of the different customs are compared to each other and a rough ranking of their supposed reliability is given.

As already mentioned in the first chapter, Seabury uses customs data from six different nations. An idea of the quality of those datasets would give extra insight in sudden movements. There is however no good method to validate the different dataset. The real trade volumes are not issued, so a strict accuracy test cannot be conducted. The goal of this section is to describe trends and movements in the CR's series and based on their consistency, a very rough ranking is created. To our knowledge, this is the first assessment of the different custom data sources based on their container ratio.

In Section 4.4.1 the CR's of the different sources are compared to each other from a high perspective. The total trend lines are interpreted and some general findings are listed. In Section 4.4.2 a categorical comparison is made between the different datasets, based on selected G1 groups.

4.4.1 Total

In this section the total import and export CR's are compared to each other. The customs listed below all dispose CR's. This is in contrary to the majority of the customs, who only name one trade volume. The weighted CR's described in Chapter 2 are used.



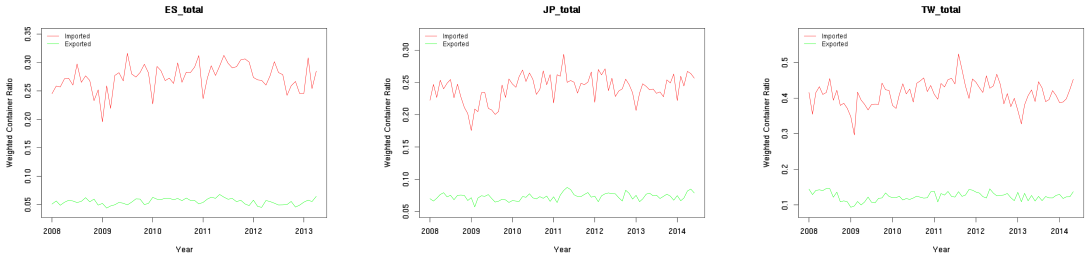


Figure 17: The global import and export CR's for all the considered data sources.

Figure 17 denotes the CR's for the different dataset. The timelines of the datasets are slightly different. The first general note is that CR's of the imported flows are higher than the exported CR's. This is due to the sample taken; all are developed economies with high consumer consumption bases. This results in relatively more containers with finished products imported than exported. In the EU sample, the two series move more or less parallel to each other.

The development seen in the US mentioned earlier in which the CR of the exporting trades is higher than its imported side, cannot be observed in the other samples. This is an indication that a change in behaviour of a nations economy has a significant effect on its CR.

Moreover, in the UK there seems to be a huge drop in 2008. This could be caused by a different calculation methodology and/or definitions. The underlying export and import quantities are more or less stationary, eliminating changing trade volumes as a cause. The drop is further analysed in the next section. Furthermore Figure 17 denotes comparable series for the Japan, Taiwan and Spain sample. On a global level their series seem to move in the same fashion. This could be a first indication of consistency among those datasets. A more general note could be made concerning the levels and trends of the datasets. Again, the volatility and fluctuations of the CR's are striking. This can be interpreted as the poor explanatory power that time, on its own, has. It is expected that on a deeper level, some of this variation will be removed.

4.4.2 G1 Level

In this section the datasets are compared to each other on a G1 level. This will give an indication about how certain commodities are shipped that belong to the same category. Although reliability cannot be determined solely based on these CR's, consistency of series does indicate a trust worthier image. The selected G1 groups are G (Raw materials, industrial consumables and foods) because of their high absolute weights, and the L group (Consumer Fashion Goods) because of its expected high CR.

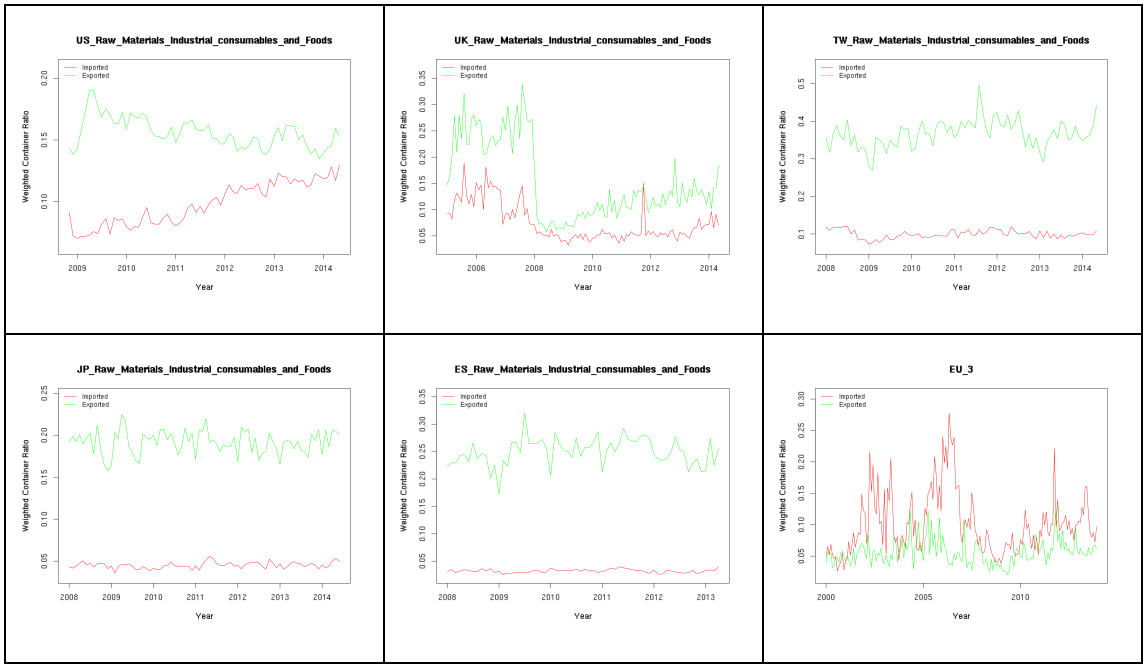


Figure 18: The import and export CR's of the different datasets for the raw materials, industrial consumables and foods G1 group.

In Figure 18 the CR's of the G-group are denoted for the different datasets. The Eurostat data, denoted by EU, does not have the mentioned G-coding. It uses the so-called NSTR coding and the therefor its categories need to be translated for direct comparison grounds.

The trends and levels are somewhat similar to the behaviour seen of a level up, denoted in Figure 17. This is caused by the enormous weights involved in the G group. This puts large factors on the observations, which even can be seen on global levels. Again the Spain, Taiwan and Japan datasets seem to have the least volatility in their series. The UK dataset still contains the large drops of CR in 2008. The EU series, representing the petroleum related products, has more volatility in the series. There seems to be an effect of outliers, which cannot be neglected.

These results could give an indication of poor classification that Seabury currently uses. Especially in this class, the variation of the cargo is too big. While lumberjacks are sometimes shipped in containers, woodchips are a familiar dry bulk commodity shipped in ship vessels. The types of industries present in a country have effect on the CR at this level. To give more substantiated statements concerning the reliability of data sources based on this class, the CR's need to be explored at least at G4. This is however out of scope in this report.

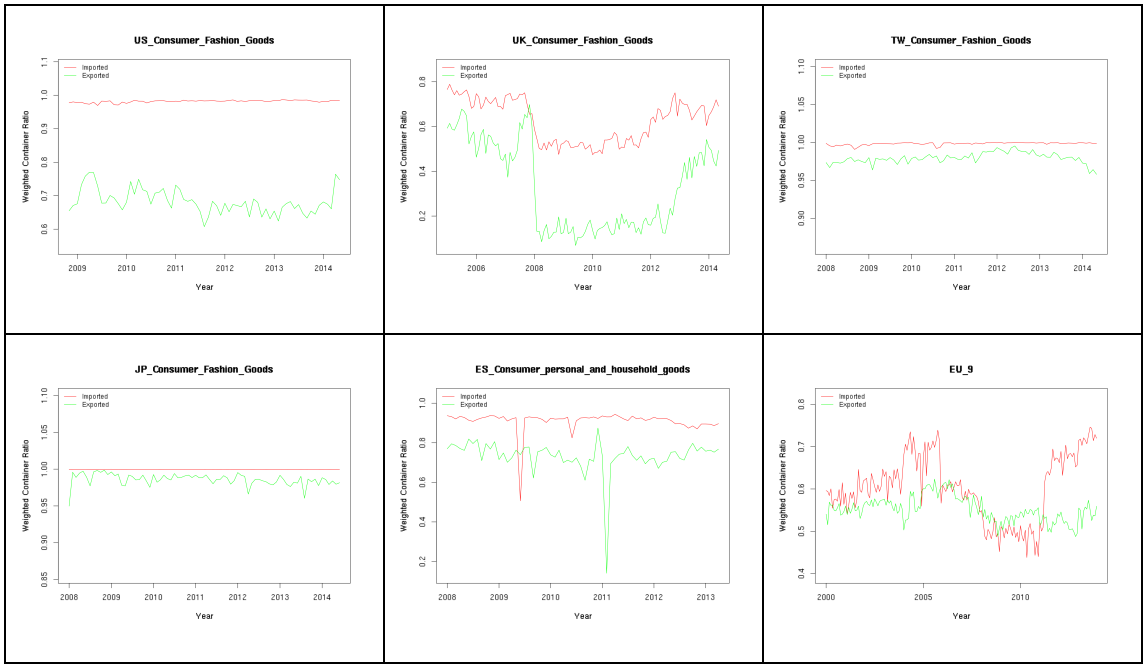


Figure 19: The import and export CR's of the different datasets for the consumer fashion goods G1 group.

The most informative comparison is denoted in Figure 19. Here the Consumer Good Fashion groups are compared to each other. The NSTR mapping of the Eurostat data, is not very specific on consumer products (Tavasszy et al., 2011). Products like machinery, transport equipment manufactures articles and miscellaneous articles are all grouped, although the internal CR's are different. The L group is a group they should contain CR's of close to one. The underlying reason for the low exporting CR of the US, was shown in Chapter 3. The low CR's of the UK and ES dataset is striking, especially on the importing side. The trade volumes of Spain are again small, giving not representative single shipments a (too) high weight. The Taiwan and Japan datasets however, show high CR's for both the exporting and importing trade lane. The effect of low CR's subgroups is small, leading to high CR's as expected. The amount of variation through the years is small, giving a steady and even container ratio. Also other G1 groups with traditional high CR's (like the Live Animals and Consumer Products) have denoted trust worthy and consistent trends and levels for the Japan and Taiwan datasets. This leads to the statement that, based on the categorical analyses of the this section, Taiwan and Japan denote the most stable and reliable CR series.

The aggregations on a regional level for the different datasets are not shown, but placed in the Appendix. The resulting graphs do not have explanatory power, just like in the previous section/ It is expected that some coherence exists when looking at the location of import and export. Hence, just like the regional analyses of Section 3.3, there is a lot of variation present among all the figures. The type of product shipped is again related to the considered region, which makes unbiased observations complex. Therefore, the explanatory capabilities of the regions are limited, given the large variation.

4.5 Import/export cross dataset comparison

In this section the reliability of the different sources are tested in another way. In a trade lane, volume is recorded in two places; the exporting country under exports and the importing country under imports. Under normal circumstances these two volumes should denote the same numbers; hence the CR's series should lay on top of each other. If a deficit exists, at least one of the sources is reporting in a different way. It is impossible to validate the series since the "real" trade numbers are not given out, but by comparing the sources against multiple other sources one can create an intuition regarding the reliability. Again, the reliability is determined based on consistency. Given the aggregation in the Eurostat data on country source level, it is not feasible to perform a similar analysis on this dataset.

Figure 20 denotes the results of this analyses. The figures show two time series; one denoting the exporting CR to that specific source and vice versa. Some remarkable observations can be made, as explained below.

In Figure 20 (a) the results of the comparison of US to Japan are denoted. So the exported CR of the US to Japan is compared to the imported CR of JP from the US, and vice versa. The trade lane US to Japan shows more or less identical CR's through time. The trade lane Japan to US on the other hand, has a level shift throughout the years. Hence, Japanese export and US import do not report the same trade volumes/ratios. In Figure 20 (b) the trade lane Taiwan and US is displayed. A difference between the two sources is denoted for both directions again involving the US import. The US reports higher CR's compared to Taiwan. This can directly be coupled to low import CR's of the US, denoted in Section 5.1. The problem the US has with the labelling of containers is not apparent when Taiwan is exporting its containers. This is an extra argument for the labelling deficit of the US mentioned earlier. Note that this deficit is not apparent on the exporting trade lanes. The CR's of both directions are more or less equal.

Figure 20 (c) the trade lane between Japan and Taiwan is displayed. Compared to the others, the CR's in this trade lane can be labelled as identical. For both directions, the reported CR's lay on of each other throughout time. The trade lanes are substantial and also on deeper (G1) levels the CR's are identical. Hence, the reporting mechanisms of Taiwan and Japan are consistent regarding the CR's. This is in line with findings in Section 5.2, where Japan and Taiwan have denoted high CR's for categories that are rather containerized. Figure 20 (d) denotes the trade lane of the UK and JP. Again, differences are observed between the reported CR's. The UK data shows more stable series with higher CR's compared with Japan. The differences are apparent for both directions, however UK export to Japan import

Figure 20: This table denotes the trade lane comparisons for the different data sources. FLTR the figures denotes: (a) Trade lane comparison between US and JP, (b) Trade lane comparison between US and TW, (c) Trade lane comparison between JP and TW, (d) Trade lane

copes with a large difference. The difference is remarkable and could be caused by the consistency of the UK dataset, discussed in Section 5.2. Figure 20 (e) denotes the reported CR's for the trade lane between Spain and Japan. This is a relative small trade lane, hence the effect of single shipment/outlier cannot be ignored. The supposed effect of these entries are especially apparent in the direction Spain to Japan. Ignoring those spikes, the CR's do show the same path. In the other direction, a level shift can be denoted. The differences seem to get smaller through time, indicating that reporting is are more identical. Figure 20 (f) denotes the trade lane UK and US. The previous figures indicated that both the UK and the US issue higher CR's than other sources. This deficit is small when comparing the US to UK trade lane. The two series follow each other closely. In the other directions however, a bigger difference can be denoted. The US importing trades seem to be reported in a different way than UK, not for the first time. Figure 20 (g) denotes the CR's between the UK and Taiwan. It shows comparable series to the trade lane US and Taiwan; in the direction UK to Taiwan the difference between the CR's is bigger than in the direction Taiwan to the UK. This implies that there is that reported UK exports and TW imports are fairly different. The UK export CR's have shown flaws in all the above figures and therefore its reliability can be questioned. Figure 20 (h) denotes the CR's of the trade lane US and Spain. The two sources report the CR's rather similar, given that the series follow each other nicely. Even the US import does not display the same trends observed in previous analyses. However, a small structural difference does exist and one should consider the relative small size of the trade lane.

Finally, Figure 20 (i) denotes the trade lane between Spain and Taiwan. The series are comparable to the CR's of the trade lane between Spain and Japan. Again, this is a relative small trade lane, hence the effect of single shipment/outlier cannot be ignored. From Spain to Taiwan the CR's are close to one. At some places in series the lower CR's, denoted by spikes, show the effect of outliers. In the other direction, a level shift can be denoted. The differences seem to get smaller through time, indicating that reporting is more identical.

All with all, the above analyses have lead to some interesting findings. The reliability of the sources cannot be determined by a cross-comparison of the import and export trade flows, given the lack of an absolute validation sample. However, if it is assumed that consistency is an indicator for trustworthy results, the sources can be rated.

The datasets of Taiwan and Japan have shown similar movement concerning the CR's through time. Trade flows that are reported in those two different sources, show identical

numbers. The US importing trade flows show significant differences compared to most of the other sources. The exporting side is more consistent, especially with Spain and the UK. Furthermore, the UK and Spain have denoted both similar and different results.

When combining these results with earlier findings, the parallels can be drawn. In the above analyses Taiwan and Japan showed promising results. This is in line with the findings of Section 4.2 in which the Eastern countries have shown consistent and explainable results for the G1 categories. Again, it is possible that both datasets show structural errors and that another dataset displays categorical CR's and trade lanes correct. However, the proven consistency is striking and one could argue that it is an indicator for reliability. The other datasets have showed mixed results. Therefor, making substantiated statement concerning their performance is not possible. However, the US dataset is extremely large and although the trustworthy of some of its CR's could be discussed, its size does imply solid CR's on a categorical basis (e.g. for G4 levels).

4.6 Relationship between CR and weight

Not only the history of the series, but also the month of the year and the type of product shipped have an effect on the probability that volume is shipped by containers. In this chapter an analysis is conducted on the relationship of CR and aggregated traded weight. The objective of this section is to show how the CR metric is able to capture underlying trends of trade data and thereby giving new insights in the data.

The background of this relation is already discussed in Chapter 2. The United States historically exports a lot of bulk material and imports goods that are shipped by containers. This mismatch is called the "Trade Imbalance" and results in a large number of empty containers on US soil (Robinson, 2007). As mentioned, a solution cloud be using containers to ship bulk material. Although this might been not economic feasible in the past, the containers available at a rather high discount. It is expected that transport companies supply container shipments for low freight rates to ship bulk material out of the country (Gurning, 2007). Hence, the hypothesis is formed that on average, the Container Ratio decreases if the weight of the shipment increases for certain groups. This expectation is tested using the different data sources.

The above statement is tested using the provided datasets. Due to the large number of observations, fitting a line on the data points using ordinary least squares techniques is not realistic. OLS tries to create a line of which the squared distance is minimal. When considering more than a million observations, this minimization step is not feasible. This research focuses on classic bulk materials which are shipped by bulk carriers but considering all materials is not feasible. Two G2 groups are selected: the "GA" group representing the "Foodstuffs & Beverages for human consumption" and the "FF" group representing "Other chemicals". However, only the dry bulk category is discussed in this section.

First of all, the regression is performed on the export and import groups for the total G1-group "Raw materials, industrial consumables and foods". Performing a short analysis on the higher level, will help in the evaluation of consistency through the different levels. The results are denoted in Table 3. It is denoted that for both the import and export trade flows there exists a significant estimated coefficient. Both of the coefficients are negative, indeed drawing a correlation between an increase of the shipped weight and the smaller use of containers. The coefficient can be interpreted as follows: for every ten million tons of cargo imported extra, the CR decreases with 0.0222 ceteris paribus. Also on the export regression, the same statement can be made. While this relationship may not seem very big, note that the volumes shipped do differ on a large scale.

Country	Estimated Coefficient	Standard Error	t-value	P(>|t|)	R-squared
US Export	-1.542e-09	2.340e-10	-6.591	9.1e-09 ***	0.4006
US Import	-2.218e-09	2.679e-10	-8.278	9.35e-12 ***	0.5132

Table 4: Results of the regression on the G-group.

Both of the regressions are plotted in Figure 21. Although the R-squared of both fits is not high, a large portion of the variation can be explained by means of the regressions.

This relationship is explored using a standard OLS regression. It is tested whether including an exponential term or power would increase the performance of the regression significantly, this was not the case. It is also illustrated in Figure 21, indicating that there is no sign that a non-linear regression would increase performance. The p-value for the t-statistic when checking for significance of the estimator (hence the hypothesis of b=0) can be rejected. Both estimators are significant. The import side seems to more affected by weight changes than the export side. However the difference is small; the standard errors only vary 14%. Hence in the G group, an increase of total shipped material is correlated to the weighted CR.

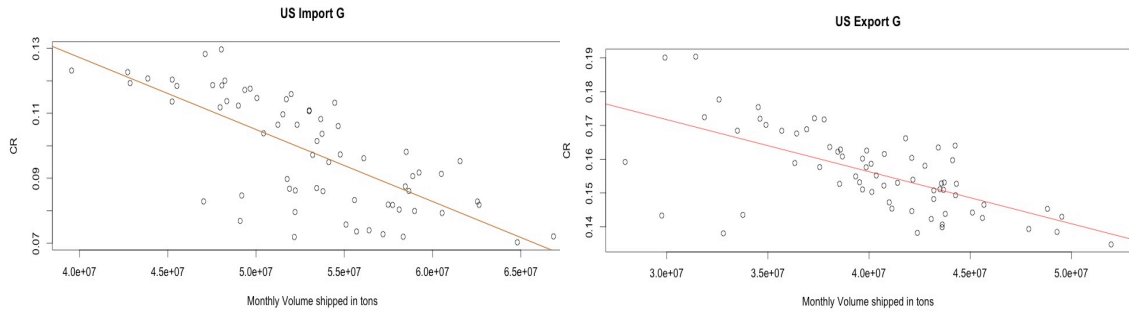


Figure 21: The regression lines of the Import and Export CR against the total shipped weight in the US.

All the regressions are also checked for heteroskedasticy and endogenity, using the summary function build-in in R. An example check can be seen in Figure 22. It denotes the summary statistics of the residual analysis of the export regression of the US. The plots illustrate whether the two most important of the OLS are violated: does there exist serial correlation among the residuals and is the variance equal throughout the regression. Especially the

second plot, which shows how the residuals are plotted against their theoretical quantities, is an indicator that the residuals are indeed normally distributed.

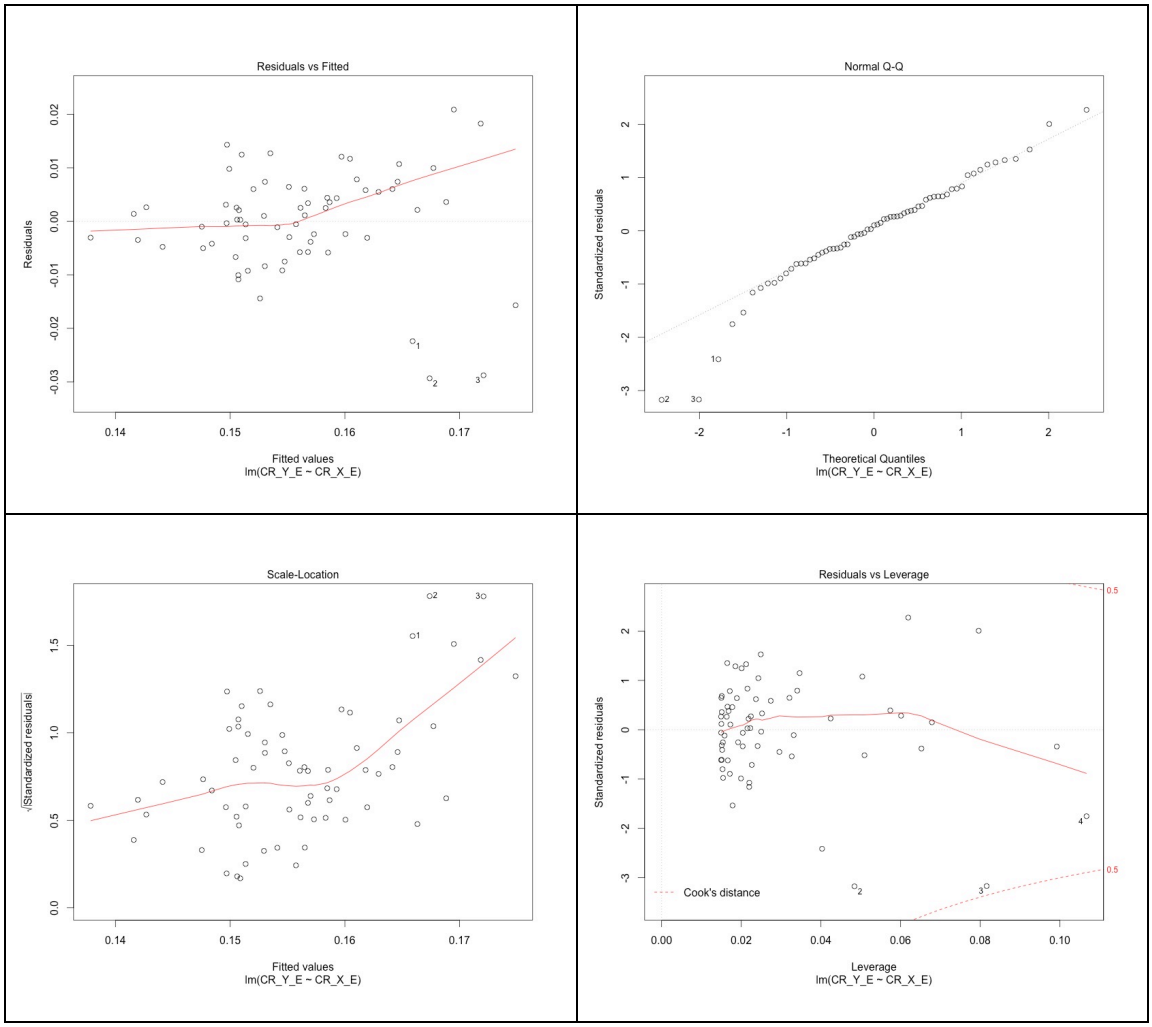


Figure 22: Detailed information concerning the residuals of the above export US regression.

4.6.1 Foodstuffs and beverages for human consumption

The next step is to conduct the same analyses on the G2 groups. In the previous section it was showed that there exists a correlations between the weight of the trade lane and the CR. This is now tested on the "GA" group, representing "Foodstuffs and Beverages for human consumption".

Country	Estimated Coefficient	Standard Error	t-value	P(>|t|)	R-squared
US Export	-1.543e-08	1.624e-09	9.502	6.54e-14 ***	0.5814
US Import	-1.689e-07	-2.340e-08	-7.217	7.17e-10 ***	0.4448
UK Export	-1.372e-06	3.778e-07	-3.632	0.000427 ***	0.1062
UK Import	-1.971e-08	1.027e-08	-1.919	0.0588	0.0462
ES Export	-5.703e-07	2.307e-07	-2.472	0.0162 *	0.0897
ES Import	-1.185e-07	8.664e-09	-13.68	<2e-16 ***	0.7512
TW Export	-1.631e-07	3.763e-07	-0.434	0.666	0.0025
TW Import	-4.544e-07	1.051e-07	-4.321	4.7e-05 ***	0.1993
JP Export	-7.621e-06	1.099e-06	-6.932	1.18e-09 ***	0.3873
JP Import	-1.971e-08	1.027e-08	-1.919	0.0588	0.0462
EU Export	2.281e-10	5.734e-10	0.398	0.691	0.0019
EU Import	-1.528e-09	3.972e-10	-3.847	0.00017 ***	0.0819

Table 5: This table denotes the results of regressing the average shipped weight in the GA category against the weighted CR. Significance codes: 0 '*' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '**

The results of the regression are denoted in Table 5. The regression is performed on all the six datasets, where in the Eurostat data the equivalent codes are used. The interpretation of the coefficient is similar to Table 4: for instance, the US export CR decreases with 0.0154 for every million tons exported extra, ceteris paribus.

For the US both estimated coefficient are negative and significant. Especially for the US Export, the relationship seems to be strong. The R-squared of trade flow is .58. So although the coefficient is not the highest of the list, the relationship seems to be working for the most observations. This is in line with the earlier stated hypothesis, in which the surplus of imported containers are used in months where the export is low. Again, there exists fluctuations among the datasets. The ES Import and UK Export denote strong correlations between the CR and the amount of shipped weight in this subgroup. These are trade flows that show have small volumes, it could be that this affected the significance of the regressors. Moreover, the observations do not lay close to the regression line. This is expressed in the low R-squared, indicating that other variables introduce a noise in the regression. This is in line with the findings of Section 4.3 and Section 4.4, in which the historical movement and specific category proved to be important proxy's for the CR.

In Figure 23 a selection of the significant datasets are plotted with their regression lines. The regression of Japan denotes to have residuals that are not normally distributed. The residuals of the regression showed serial correlation and there for endogeniety is present. This is an indication that, although the p-value and R-squared denote promising numbers, the regression is not correct. The same results was found for Taiwan import. Hence, although the Asian datasets have showed significant relationships, the regressions fail other technical demands. This indicate that the relationships cannot be trusted.

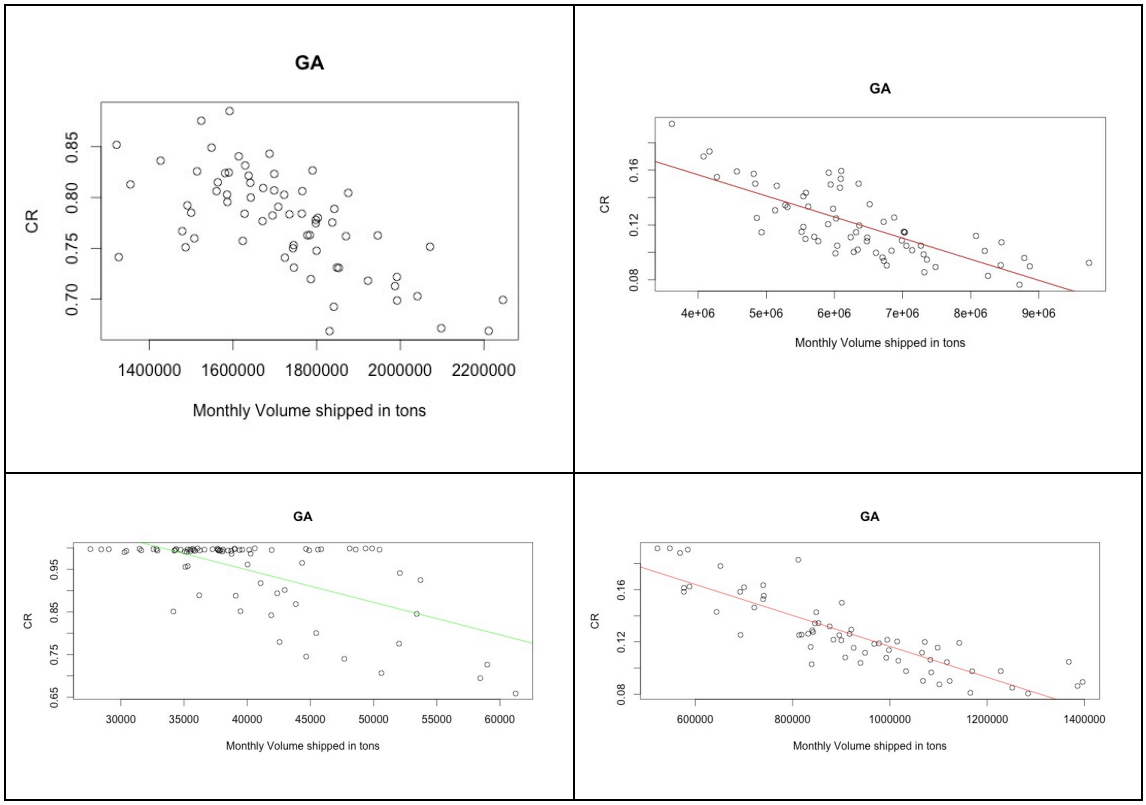


Figure 23: Relationship between CR and monthly weight in GA group and their regressed lines. Top-left Import US, Top-right Export US, Bottom-left Export Japan and bottom-right ES import.

Based on the analyses conducted in this chapter, some findings can be listed. In this chapter the assumption that there exists a direct relationship between the monthly average shipped weight and its CR was tested. The research was limited to the G, GA and FF group. (For the FF group results, consider Appendix B.4.)

There seemed to be little consistency among the datasets. The G and GA group in general showed a significant effect. Especially the US dataset has denoted a relationship where a smaller shipped weight results in higher CR's. However, the results of the other countries were more volatile. Within the FF group only Spain, Taiwan export and Japan export showed significant coefficients. For the other datasets, the linear regression was rejected.

A general remark could be made regarding the level of the categories of the analyses. It was chosen to perform the analyses on G2 level. However, within the G2 levels, especially in G groups, the type of products and hence the CR's are different within groups, as shown in Table 3. It could therefore be interesting to observe how the correlation exists on deeper levels. This would remove some of the bias created by independent moving CR's of subgroups. Moreover, in Chapter 3 it was mentioned that the data set was removed from traffic. However, for this specific application this traffic has information concerning the (empty) container reverse logistics. When measuring if the decrease of empty containers in a trade lane is correlated with the increase of bulk filled containers, one would gain more certainty on the made statements.

The main objective of this section was to illustrate the predictive capabilities that the metric CR could have to predict container flows. Although the presented relationships could be tracked using the absolute numbers eventually, the CR could be used for trend spotting rather easily. The findings of this chapter could be used in the future for predicting CR's. A model could be formed in which certain commodities, the historical trend and type of data sources determine a prior of the CR as showed in this chapter. However, also the size of the volume could function as a proxy influencing the CR.

5. Conclusion and Discussion

In this report, an explanatory study was conducted on so-called container ratios. A container ratio (CR) denotes the distribution of commodities shipped in containers, against commodities shipped outside of a container. Only a hand full of international customs issue container specific information, namely the UK, US, Taiwan, Japan, Spain and an aggregate of European counties. A large amount of transport companies however, are interested in this information for the other countries, which do not issue any information, as well. Before an assessment could be conducted, problems associated with the data format were overcome. When assessing the quality of the different sources, the results were mixed. Due to a lack of validation points, a direct comparison between the sources was conducted. Taiwan and Japan showed more consistent results compared to the other sources. However, given the volatility of the time series, more research is necessary. This information could be used in the next phase of the project, where a model will be built estimating the CR's for the mentioned countries.

First of all, the advantages of using a container ratio over absolute numbers were discussed. To our knowledge, no scientific study has been conducted specifically on the ratio of containers as a metric to assess datasets. However, there are a number of arguments for using the CR in the proposed context. First of all, the CR is able to visualize relative trends in containerization. It is able to isolate fluctuations in trade lanes and denote how containerization behaves for subgroups. When looking at the absolute trade volumes without considering the total volume of the trade lane for that cargo, the resulting insight could be rather biased. Another reason revolves around one of the goals formulated in this report. One of the main objectives was to assess the quality of the different sources of the dataset. By aggregating CR's on groups that have known containerizations, one is able to assess the quality of that source. The reliability of sources issuing unexpected CR's could be questioned.

One of the main objectives of the report was to come up with a methodology that was able to handle the data size and complexity. This was developed in Chapter 3. Since the statistical software R originally puts the whole dataset in its active memory, several solutions were presented. In the end, the dataset was aggregated upon its country source. Only the active parts of the dataset were loaded into the main memory using SQL, while keeping track of metadata of the whole object, also the non-active parts. Moreover, a number of operations are executed to reduce data complexity. A weighing factor is deployed for the issued CR's, for aggregation purposes the Split-Apply-Combine structure was used, the labelling was made

uniform and the dataset was cleaned. In the end, the proposed scheme led to a fast and lean implementation of the dataset, as shown in Chapter 4.

After the data complexity was reduced, the influence of different variables on the CR could be explored. The category of the shipped product proved to be a good proxy for the CR, both on a high level (G1) as on a more specific level. Moreover, the historical movement of a CR was an effect that could not be neglected. When shifting to deeper categorical levels, classes with high weights had a high influence. It was even possible that a very specific G4 trade flow with a substantial weight, lead to a bias throughout the whole G1 category, as showed for the fashion group. For some G2 groups, the effect of seasonality was substantial. There seemed to be little grouping possible within the regional variable, based on the CR. It was proven that there exists a high correlation between the type of commodity and its destination. Hence, the effect of the regions was rather biased by that observation and lead to the assumption that the added explanatory capabilities of the regional variables on a group level were small; one would have to explore specific G4 codes (or even deeper) to find regional differences.

Next, the different datasets were compared to each other and where possible, an ordering of the sources was formed. Hence, the qualities of the different data sources were assessed. First the historical CR's were analysed for different subgroups. On a high global level, the volatility and fluctuations of the CR's were striking. When comparing the CR's of similar categories for different datasets, Taiwan and Japan denoted the most stable and reliable CR series. The trade volumes of Spain were small, giving unrepresentative single shipments a (too) high weight. The UK dataset was highly volatile and the US dataset showed relative low CR's for categories that should denote one, especially on the exporting side. This could indicate a problem with labelling. The type of product shipped was again related to the considered region, which made unbiased observations complex. A second method to test the reliability of the different data sources was by comparing the importing trade flows from dataset A to dataset B, to the exporting trade flows of dataset B to dataset A. The reliability of the sources cannot be determined by a cross-comparison of the import and export trade flows, given the lack of an absolute validation sample. However, if it was assumed that consistency is an indicator for trustworthy results, the sources can be rated. The datasets of Taiwan and Japan have shown similar movement concerning the CR's through time. Trade flows that were reported in those two different sources, showed identical numbers. When combining these results with earlier findings, the parallels could be drawn. The Eastern countries have shown consistent and explainable results for the G1 categories. The proven consistency was striking and one could argue that it is an indicator for reliability. The other datasets have shown mixed results. Therefor, making substantiated statement concerning their performance is not possible. However, the US dataset is extremely large and although

the trustworthy of some of its CR's could be discussed, especially on the importing flows, its size does imply solid CR's on a categorical basis (e.g. for G4 levels). For more certainty around the reliability in general, more research is necessary.

Finally, it was tested whether the CR depended on the aggregated monthly-shipped volume for selected bulk materials for all the datasets. This illustrated how a CR could be used to detect underlying trends in trade data, which would be hard to discover using traditional trade numbers. The selected groups were the G1-category G representing "Raw materials, industrial manufacturers an other supplements" and the G2 groups GA "Foodstuffs and beverages for human consumptions" and FF denoting "Other chemicals". To measure the relationship, a regression was conducted in which the significance of transported weight on the CR was tested. The results of the regressions were mixed. The G and GA group in general showed a significant effect. Especially the US dataset has denoted a relationship where a smaller shipped weight results in higher CR's. However, the results of the other countries were more volatile. Within the FF group only Spain, Taiwan export and Japan export showed significant coefficients. For the other datasets, the linear regression was rejected. Hence, the use of CR in this context proved its added value.

Before implementing the results in a CR model, a number of subjects need to be researched further. For the loading scheme of the dataset, an aggregation based on source was performed. However, when considering a bigger time horizon or data with a higher frequency, this solution is not feasible. The use of parallel computing, especially Hadoop, cannot be neglected and an implementation needs to be researched. Furthermore, most of the analyses were conducted on a G2 level. However, as shown in Chapter 3, there seems to be little consistency between the G2 and G1 group. Furthermore, trade flows with a high weight seem to have a large influence on higher CR's. It is therefore be interesting to see how CR's are being influenced on a G4 level. By doing so, the bias of high volume trade lanes is minimized leading to more meaningful and consistent CR's. Moreover, one of the objectives of this study was to form a ranking of the quality of the different datasets. Since no objective information exist on these trades, the ranking mentioned in Chapter 4 was not validated. This validation is essential before using the results of this report in a model. One could for instance opt for retrieving a physical validation sample or collect expert judgements on the quality of the data. Combining these different viewpoints will lead to a more substantiated ranking of the datasets.

The findings of this research can be used as a start point for future research. A methodology to handle the dataset was introduced together with dataset specific operations. When assessing the quality of the different sources, the CR resulted in mixed observations. Due to a lack of validation points, a comparison between the sources was conducted. An intuition on

the reliability of the different sources was formed, however it is recommended to combine other methods to asses the quality of the sets. Hence, some of the findings need to be substantiated before it can be used in a model. From there on, a next phase is to construct a methodology, which is able to estimate to CR of countries that only issue limited information concerning their trade flows. Building a reliable model could ultimately lead to more efficient supply chain logistics, with all the economical and ecological advantages that this implies.

References

Adler, Nenadic, Zucchini & Glaser, 2007. The ff Package: Handling Large Data Sets in R with Memory Mapped Pages of Binary Flat Files. *UseR!*.

Brooks, E., 2012. Container vessels prepare to turn grain shipping logistics upside down. *FEED Business Worldwide*.

Chen, N., 2012. Determinants for assigning value-added logistics services to logistics centers within a supply chain configuration. *Journal of international logistics and trade*, pp.3-41.

Cormode, G. & Duffield, N., 2010. Sampling Big Data. New York, 2010. *Conference: ACM SIGKDD 2014.*

Crawley, M.J., 2013. *The R Book*. New Dehli: Wiley.

De Grace, J., 1968. *International Shipper's Atlas*. San Fransisco: Harper Group.

Diaz, R., Talley & Tulpule, 2011. Forecasting Empty Container Volumes. *The Asian Journal of Shipping and Logistics*, 27(2), pp.217-36.

Eurostat, 2014. *Transport Database*. [Online] Available at: http://epp.eurostat.ec.europa.eu/portal/page/portal/transport/data/database. [Accessed 5 September 2014].

Fleming, D.K., 1997. World container port rankings. *Maritime Policy & Management: The flagship journal of international shipping and port research*, 24(2), pp.175-81.

Fransoo, J., 2008. Ocean Container Transport: An underestimated and critical link in global supply chain performance. *Internal Report, BETA publication*, p.27.

Grossmann, H., 2007. Growth potential for maritime trade and ports in Europe. *Intereconomics*, pp.226-32.

GStat, 2010. Big Data in R. [Online], *Israel Statistical Association.* Available at: http://statistics.org.il/wp-content/uploads/2010/04/Big_Memory%20V0.pdf [Accessed 21 October 2014].

Gurning, S., 2007. An analysis of implementing container transport for wheat cargoes between Australia and Indonesia. *International Conference on Supply Chain Management and Information Systems (SCMIS),* .

Ihaka, R.&.G.R., 1996. R: A Language for Data Analysis and Graphics. *Journal of Computational and Graphical Statistics* , 5(3), pp.299-314.

Iseman, P.a., 2014. U.S. Census Bureau U.S. Bureau of Economic Analysis. *U.S. Department of Commerce*.

Jedermann, R., 2013. Sea transport of bananas in containers – Parameter identification for a temperature model. *Journal of Food Engineering*, pp.330-38.

Jones, B., 2000. Developing a standard definition of intermodal transportation. *Transportation law journal*, 27(3).

Kaluza, 2010. The complex network of global cargo ship movements. *Journal of the royal society*, pp.1092-103.

Maritime Sun, 2012. *Grain market turmoil hits depressed shipping sector*. [Online] Available at: http://www.maritimesun.com/news/grain-market-turmoil-hits-depressed-shipping-sector [Accessed 20 October 2014].

McFarlane, S. & Saul, J., 2014. *Food importers shift from dry bulk cargo ships to containers*. [Online] Available at: http://www.reuters.com/article/2014/02/14/agri-container-idUSL5N0LF3MZ20140214 [Accessed 2 October 2014].

Medhi, J., 1992. *Statistical Methods: An Introductory Text*. New Dehli: New Age International.

Nieuwsblad Transport NT, 2011. *Steeds meer bulk in container*. [Online] Available at: http://www.nieuwsbladtransport.nl/Nieuws/Article/tabid/85/ArticleID/17357/ArticleName/Steedsmeerbulkincontainer/Default.aspx [Accessed 20 October 2014].

R.3.0.1, 2008. R: A language and environment for statistical computing. *R Foundation for Statistical Computing*.

Ramanakumar, 2009. Containerization is on the fast track in India. *Journal of contemporary research in management*, pp.41-68.

Rijsenbrij, J.C. & Van Ham, J.C., 2012. *Development of containerization*. Amsterdam: IOS Press BV.

Robinson, B., 2007. Can't keep running on empty. *Cargo Systems*, pp.59-61.

Rosario, R., 2010. Taking R to the Limit: Working with large datasets,. *Los Angeles R Users' Group* .

Sabavala, A. & al-Saffar w, A., 2011. *GCC trade and investments flows: The emerging-market surge*. Economist Intelligence Unit.

Sudipto, D., 2010. Ricardo: integrating R and Hadoop. *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data*.

Tavasszy et al., 2011. A strategic network choice model for global container flows: specification, estimation and application. *Journal of Transport Geography*, pp.1163-72.

Thusoo, A., 2013. Hive – A Petabyte Scale Data Warehouse Using Hadoop. In *IEEE 29th International Conference on Data Engineering (ICDE)*. Stanford, 2013. ICDE.

UNCTAD, 2013. Review of maritime transport. In *United nations conference on trade and devolpment.*, 2013. United Nations Conference on Trade and Development.

Urbanek, S., 2013. Tackling big data with R. In *Bioconductor Developer Day*. Seattle, 2013.

Versino, C., 2010. World Trade Data: can they help safeguards verification activities? In *ANMM 51st annual meeting.*, 2010.

Versino, C., Tsukanova, M. & Cojazzi, G., 2010. An overview of the different datasets and a corresponding catalogue: Catalogue of WEB Data Services on Global Trade. *JRC Scientif and technical reports* .

Vigarie, A., 1999. From break-bulk to containers: the transformation of general cargo handling and trade. *GeoJournal*, pp.3-7.

Wickham, H., 2011. *plyr: Tools for splitting, applying and combining data v. 1.4.*. http://CRAN.R-project.org/package=plyr.

Wickham, H., 2011. The Split-Apply-Combine strategy for data analysis. *Journal of Statistical Software*.

World Customs Organization, 2012. *HS Nomenclature 2012 edition*. [Online] Available at: http://www.wcoomd.org/en/topics/nomenclature/instrument-and-tools/hs_nomenclature_2012. [Accessed 20 October 2014]

Appendices

Appendix A: Abbreviations of the G-codes

In this appendix the meaning of the G-codes mentioned in the report can be found. The first part describes the G1-codes and the following section the G2-codes.

A.1 G1-codes

G1-Code	Description
A	Live Animals
B	Temperature or Climate Control
C	Secure or Special Handling
D	High Technology
E	Land Vehicles & Parts
F	Chemicals & Products
G	Raw Materials, Industrial consumables & Foods
I	Capital Equipment & Machinery
K	Machinery parts. Components, supplies & manufactures n.e.s.
L	Consumer Fashion Goods
M	Consumer personal & household goods
Z	Waste Products

A.2 G2-codes

G2-Code	Description
AA	- Fish, Live
AB	- Poultry, Live
AC	- Other Animals, Live
BA	- Foods, Fresh
BB	- Foods, Frozen
BC	- Foods, Fresh or Frozen (not further detailed)
BD	- Foods, Cured (smoked, dried salted etc.)
BE	- Perishable Non-Foods
CA	- Valuable
CB	- Art
CC	- Dangerous Goods (DGR)
CD	- Other Special Handling
DA	- Aerospace
DB	- Semiconductors
DC	- Computers & Related
DD	- Telecommunications

DE - Nuclear Industry
DF - Radio-frequency communications, T.V., radar & navigation equipment
DG - Machinery & apparatus for scientific, medical or technical purposes
DH - Parts & accessories for scientific, medical or technical apparatus
EA - Land Vehicles
EB - Land Vehicle Parts
EC - Batched Data : Land Vehicles & Parts
FA - Pharmaceuticals
FB - Biocides
FC - Odors & Flavours
FD - Colours & Dyes
FE - Photography
FF - Other chemicals & products
GA - Foodstuffs & Beverages for human consumption
GB - Basic industrial raw materials
GC - Semi-manufactured industrial consumables
GD - Plastic & rubber industrial consumables
GE - Manufactured metal industrial consumables
GG - Consumables for civil engineering, construction & building
GH - Consumables for the textile industry
GI - Consumables for the leather & fur industry
GJ - Consumables for packaging & transportation of goods
IA - Machinery for production of physical or electrical power
IB - Machinery for agriculture, construction, mining & mechanical handling
ID - Machinery for the food processing industry
IE - Machinery for the textile and leatherworking industries
IF - Machinery for the metalworking industry
IG - Machinery for other manufacturing
IH - Machinery for general industrial uses
II - Machinery for offices, shops and similar
KA - Parts & components : power, agriculture, construction, mining, handling
KB - Parts & components of machinery for the manufacturing industries
KC - Parts & components of machinery for general industrial applications
KD - Parts & components of machinery for offices, shops, science, or techinal
KE - Parts & components for machine-tools; tools & tooling
KF - Parts & components of other machinery nes. Incl. household machinery
KG - Electrical components
KH - Supplies & consumables for the fashion industry
KI - Parts, components , supplies & consumables n.e.s.
KJ - Miscellaneous manufactures
LA - Clothing & Accessories

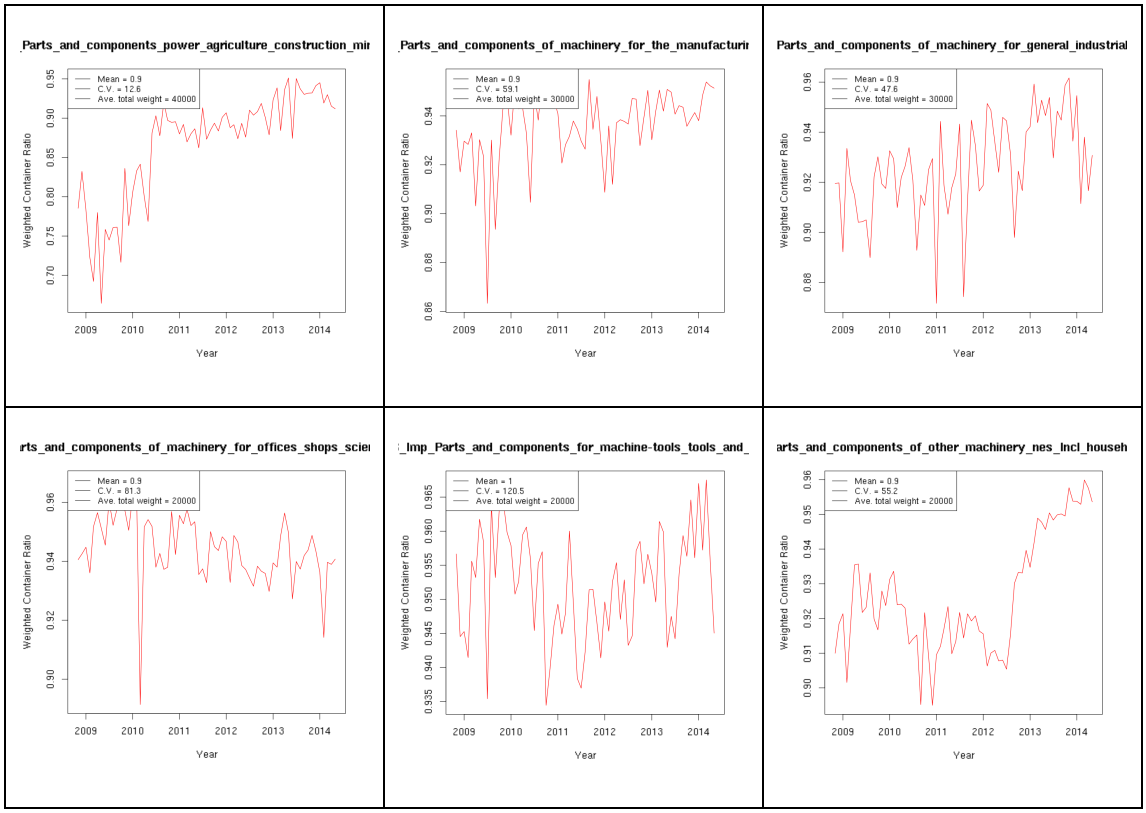
LB - Footwear
MA - Consumer goods for personal consumption
MB - Consumer goods for household consumption
ZZ - Waste Products

Appendix B: Additional results of Chapter 4

In this appendix more results of the analyses on CR's are denoted. In this Appendix, one can find results of analyses similar to Chapter 4, only peformed on other samples.

B.1 Categorial analyses US import Machinery parts group

A similar analyses as in Section 4.3.1 can be conducted on the import "Machinery parts group" K. Due to the heavy weight of non containerized machines, like tractors, the CR is rather low. Results of this analysis are denoted in Figure A.1. The weights of the groups, denoted in the top left corners of the figures, seems to be closer to each other. Weight of the trade lanes however, still has a large impact on the group level CR's. The volatility of the different trends are caused by the small trade volumes of the different lanes. Outliers have an impact on the trends of the different G2 groups, in the form of a dip or peak.



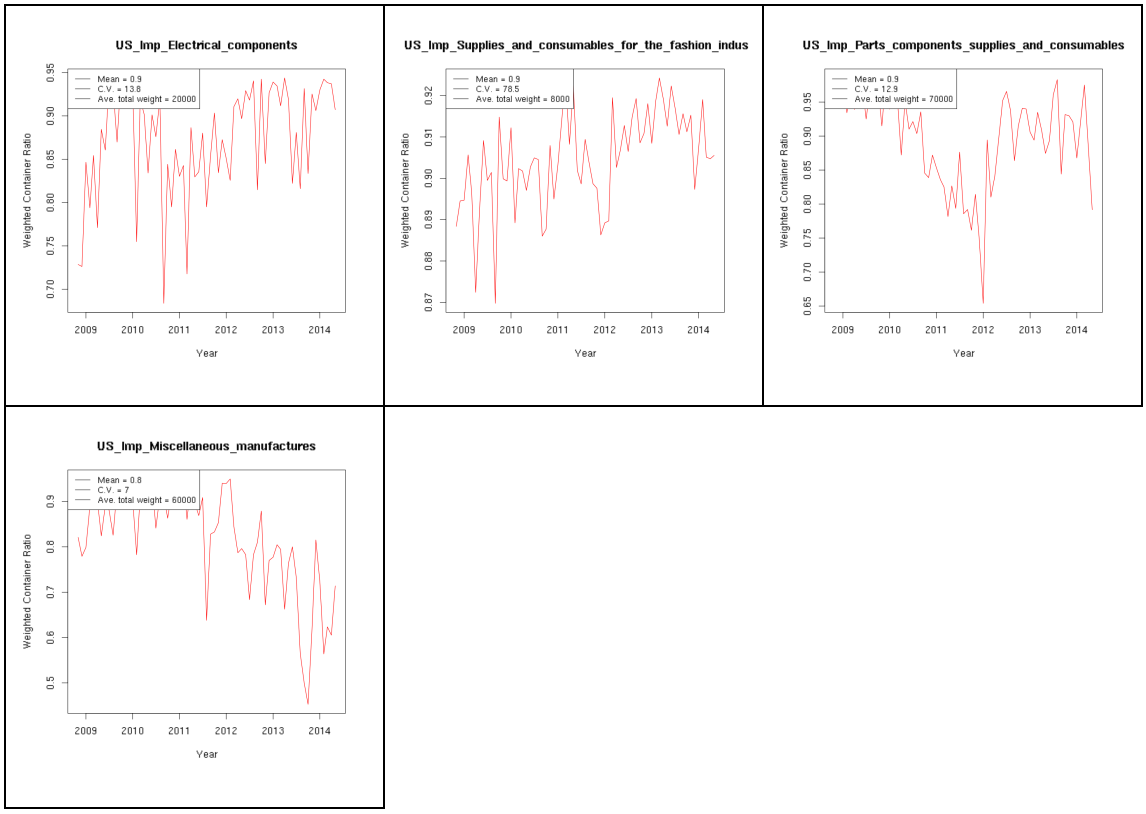
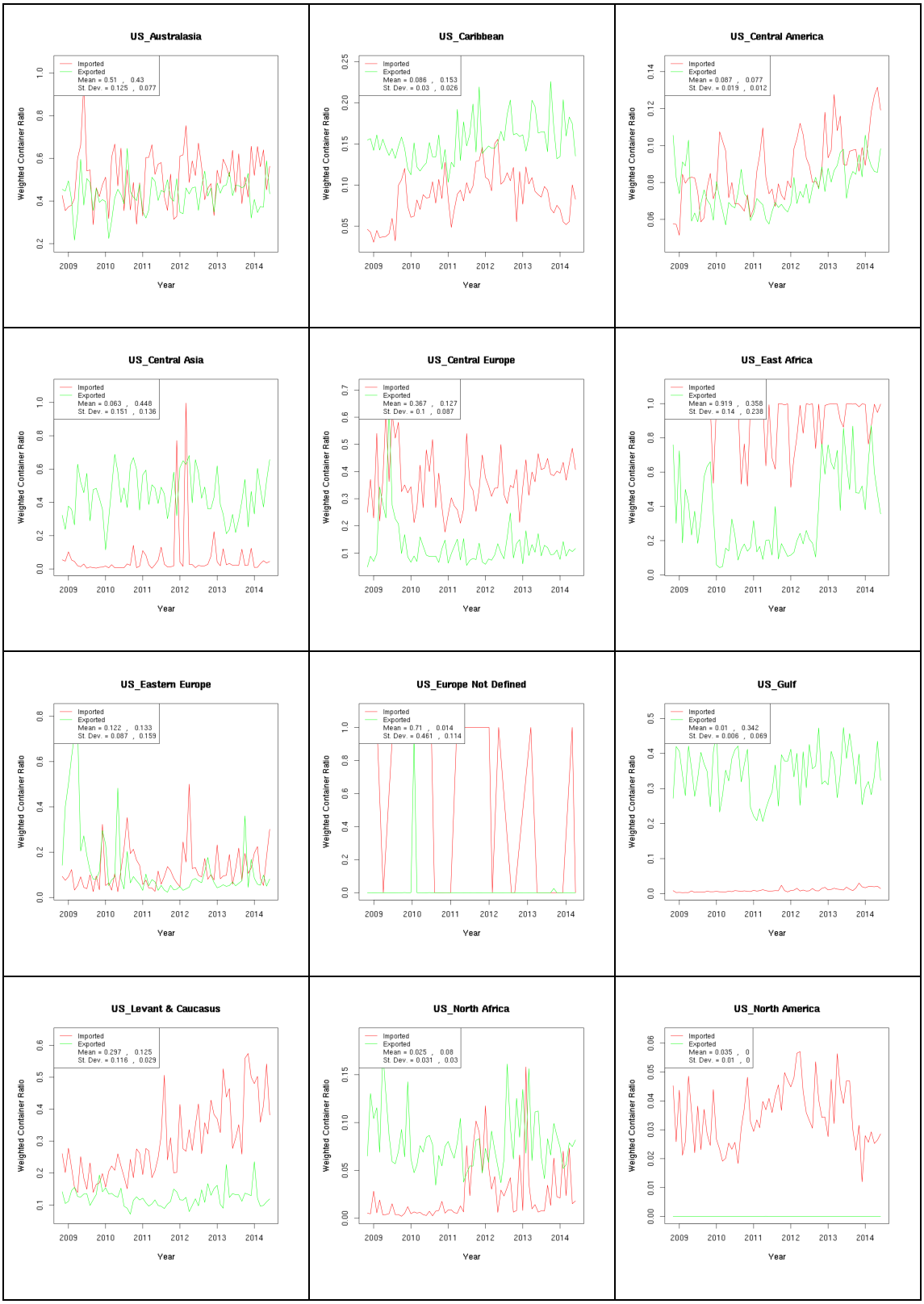


Figure B.1: The importing CR's of the G2 groups in K: Machinery parts. Components, supplies and manufactures class "K" in the United States

B.2 Results regional analyses US

In Figure B.2 the import and export lanes from and to the US are denoted. Again, the weighted CR is used and regions are defined earlier, based on the similarities among countries. On the vertical axes the weighted CR is displayed and again on the horizontal axes timelines. The legend denotes respectively mean of the imported and exported series and its variation in terms of standard deviation.


US_Australasia
Imported
Exported
Mean = 0.51 , 0.43
St. Dev. = 0.125 , 0.077
US_Caribbean
Imported
Exported
Mean = 0.086 , 0.153
St. Dev. = 0.03 , 0.026
US_Central America
Imported
Exported
Mean = 0.087 , 0.077
St. Dev. = 0.019 , 0.012
US_Central Asia
Imported
Exported
Mean = 0.063 , 0.448
St. Dev. = 0.151 , 0.136
US_Central Europe
Imported
Exported
Mean = 0.367 , 0.127
St. Dev. = 0.1 , 0.087
US_East Africa
Imported
Exported
Mean = 0.919 , 0.358
St. Dev. = 0.14 , 0.238
US_Eastern Europe
Imported
Exported
Mean = 0.122 , 0.133
St. Dev. = 0.087 , 0.159
US_Europe Not Defined
Imported
Exported
Mean = 0.71 , 0.014
St. Dev. = 0.461 , 0.114
US_Gulf
Imported
Exported
Mean = 0.01 , 0.342
St. Dev. = 0.006 , 0.069
US_Levant & Caucasus
Imported
Exported
Mean = 0.297 , 0.125
St. Dev. = 0.116 , 0.029
US_North Africa
Imported
Exported
Mean = 0.025 , 0.08
St. Dev. = 0.031 , 0.03
US_North America
Imported
Exported
Mean = 0.035 , 0
St. Dev. = 0.01 , 0
Weighted Container Ratio
Year
2009
2010
2011
2012
2013
2014

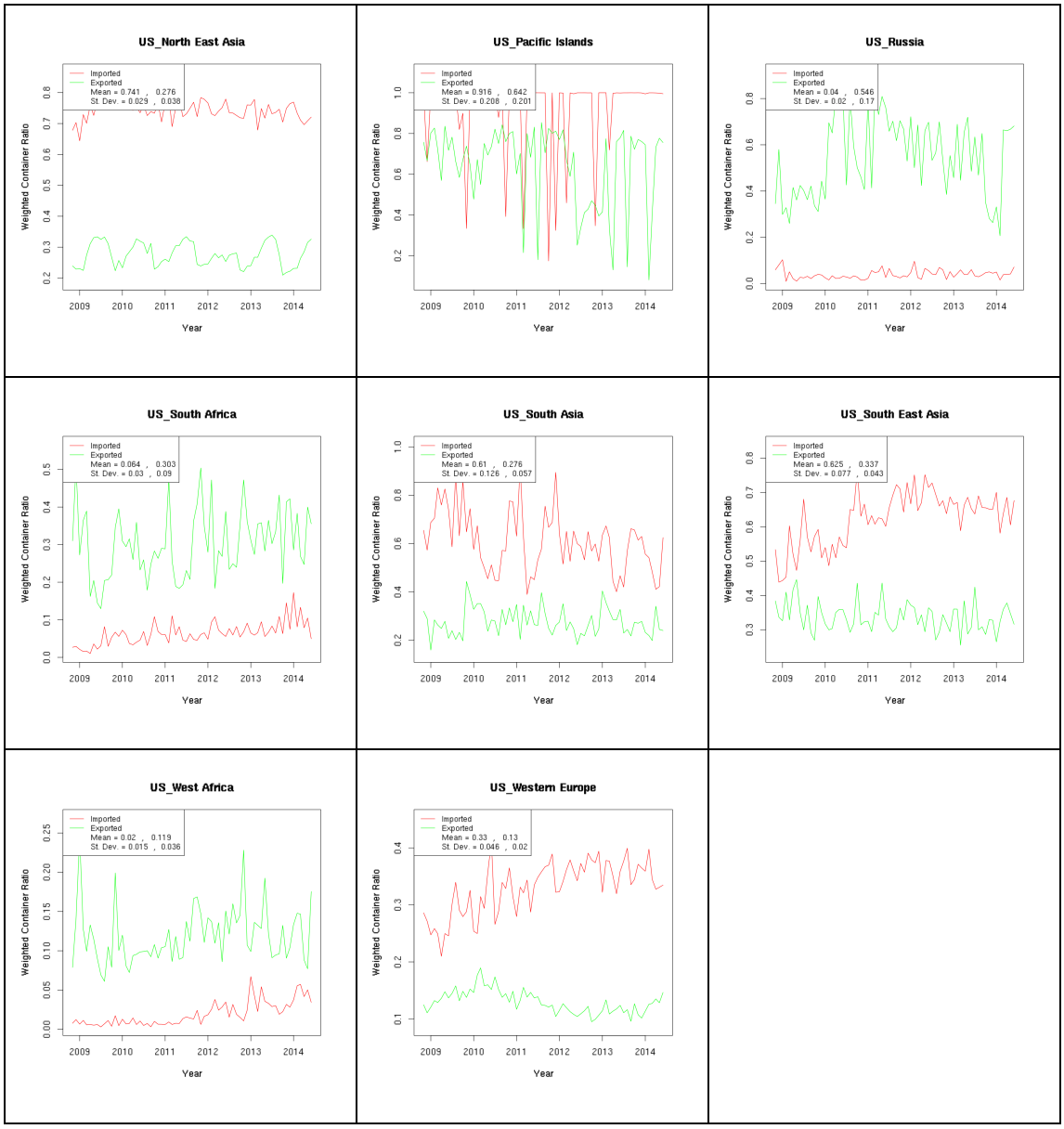


Figure B.2: Imported and exported CR's of the US split into regions

Because of the amount of figures, they will not be discussed individually. Trends among the figures will be discussed together with specific observations. Hence, there seems to be a lot of variation among the groups. The trends that are present, are highly influenced by the underlying cargo transported in that region.

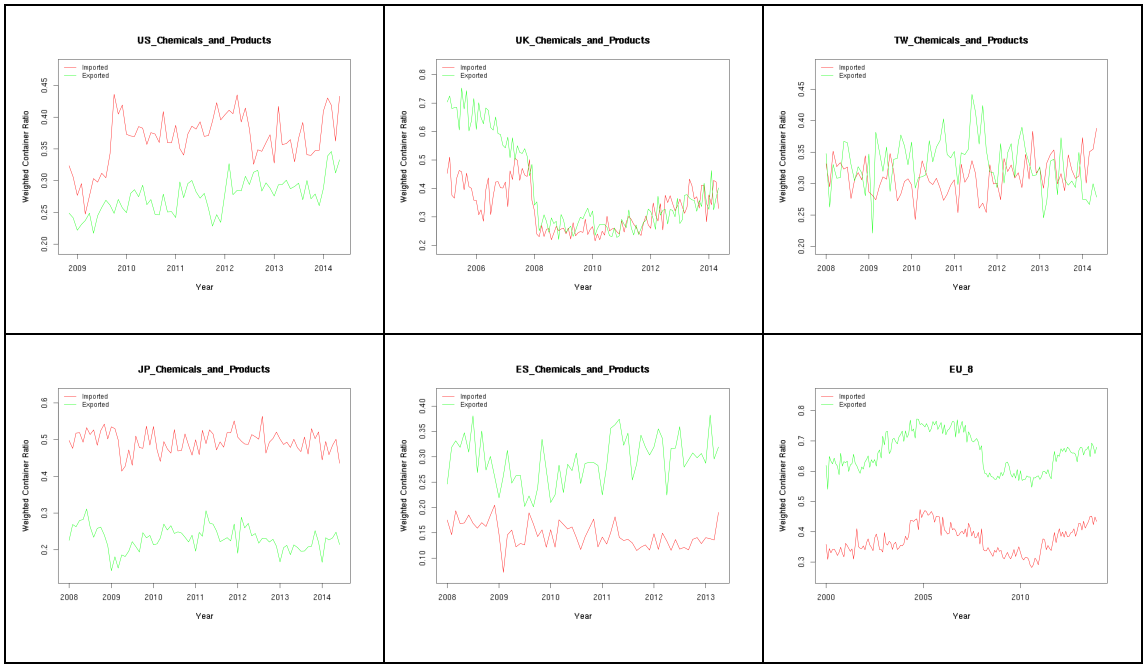
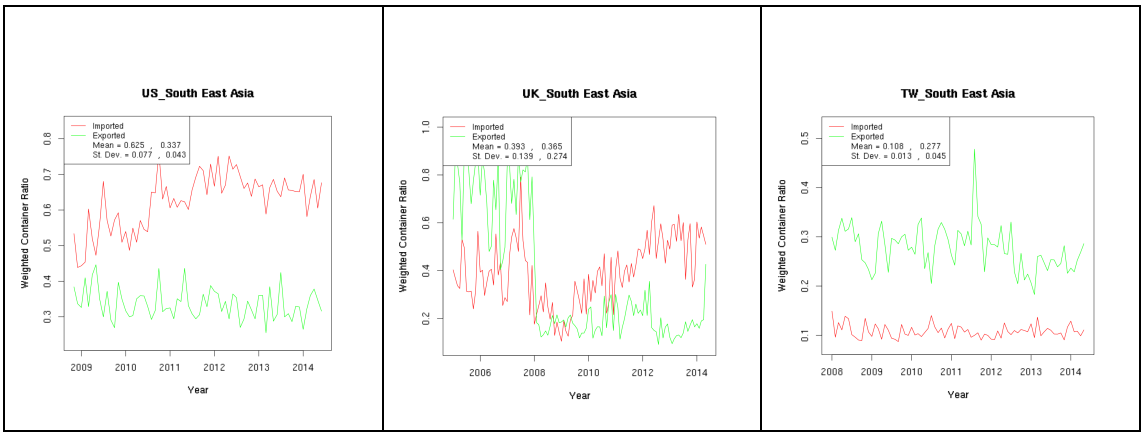


Figure B.3: The import and export CR's of the different datasets for the chemicals G1 group.

In Figure B.3 the CR's of the G1 chemicals group are denoted. The EU NSTR coding has also a Chemicals category, which makes direct comparison possible. The UK trend drop in 2008 is also present in the total EU dataset, but smaller. The inclining CR after the level shift is not present in the Spain dataset. The small trade volumes of Spain in this specific group are small, and the effect of specific shipments therefore significant. The Japan and Taiwan trend lines are again close to each other.

B.3 Regional variable compared to different data sources



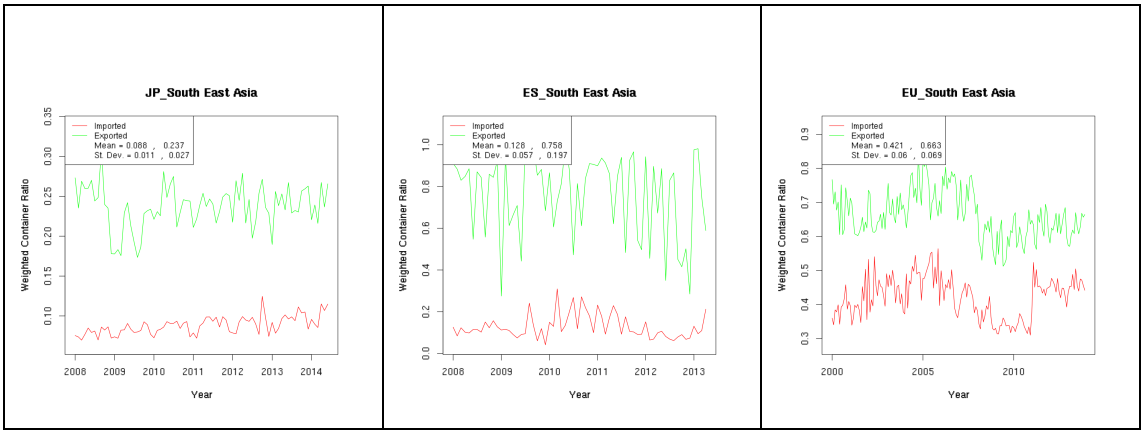


Figure B.4: The importing and exporting CR's of the different datasets to South East Asia.

There are a total of 21 regions so for feasibility grounds, a selection of those regions with interesting findings will be shown.

First of all, the container ratio's of South East Asia are analysed. The total trade volume of this region is substantial making it relevant for an in-depth analysis. The results are denoted in Figure B.4. There are again a large fluctuations in CR's throughout the data. The variation is larger compared to categorical analyses conducted in Section 4.2.

The UK shows a big drop of both ratios in 2008. This is not caused by the trade volume, but due to another factor. It might be another calculation method for CR, as mentioned earlier. The CR's are rather stables. Only in the United States and the United Kingdom show obvious trends. The exporting CR's are low throughout the different datasets, but the importing CR's are high (there is a big difference). This is in line with expectation; the region is known for its production of goods. This means that bulk material is imported by the region and containers are exported.

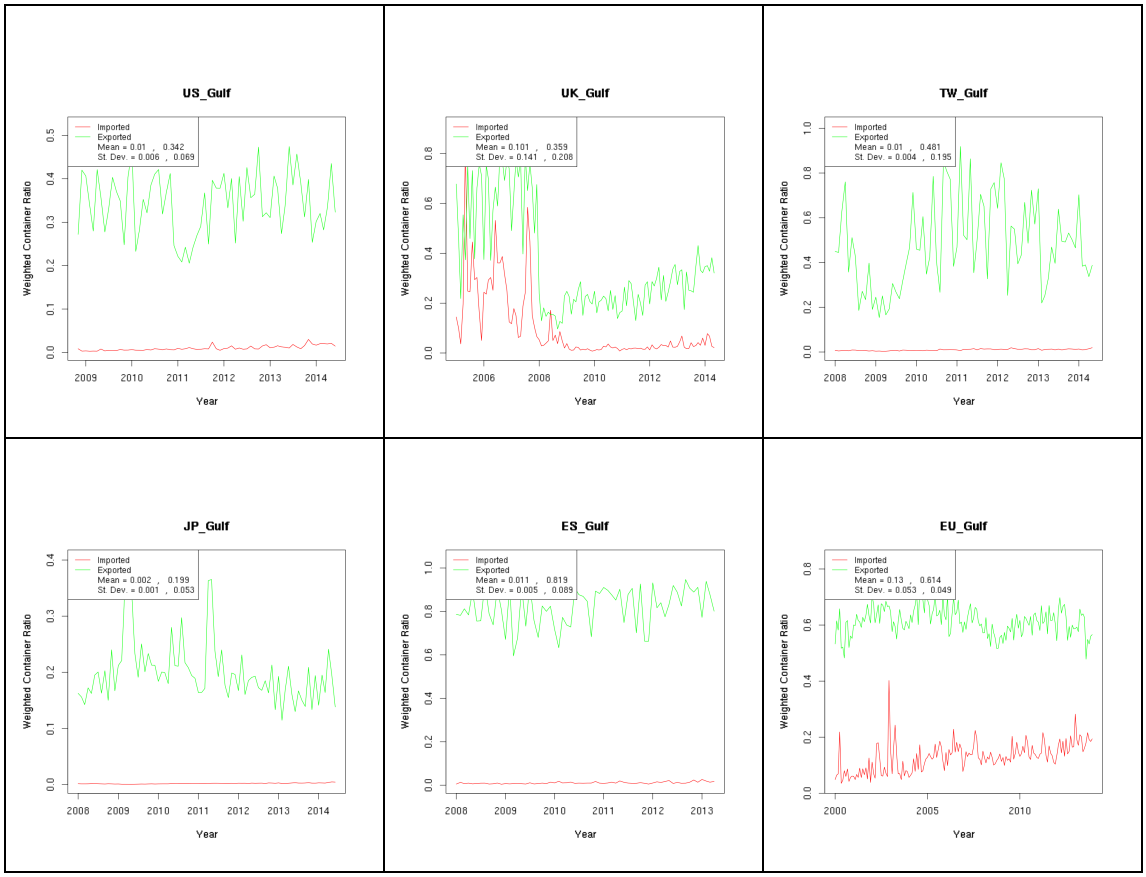


Figure B.5: The importing and exporting CR's of the different datasets to the Gulf region.

Another region that is displayed, is the Gulf region. The results are denoted in Figure B.5. The gulf region can be used as a quick reliability check, since the main export product of the Gulf region is petroleum related products (Aviva, 2011). Moreover, it has large trade volumes and have a big effect on the total CR's of above figures.

Given the large amount of exported oil out of the Gulf area, one would expect CR of close to zero on the importing side. This is the case for the United States, Taiwan, Japan and Spain. In the United Kingdom (pre 2008) and the Eurostat data however, the CR is higher than expected. After the level shift, the United Kingdom denotes more comparable levels to the other sources. This is an indication that the level shift is indeed a change in definition/labelling. The seasonality seems to have an effect on the CR.

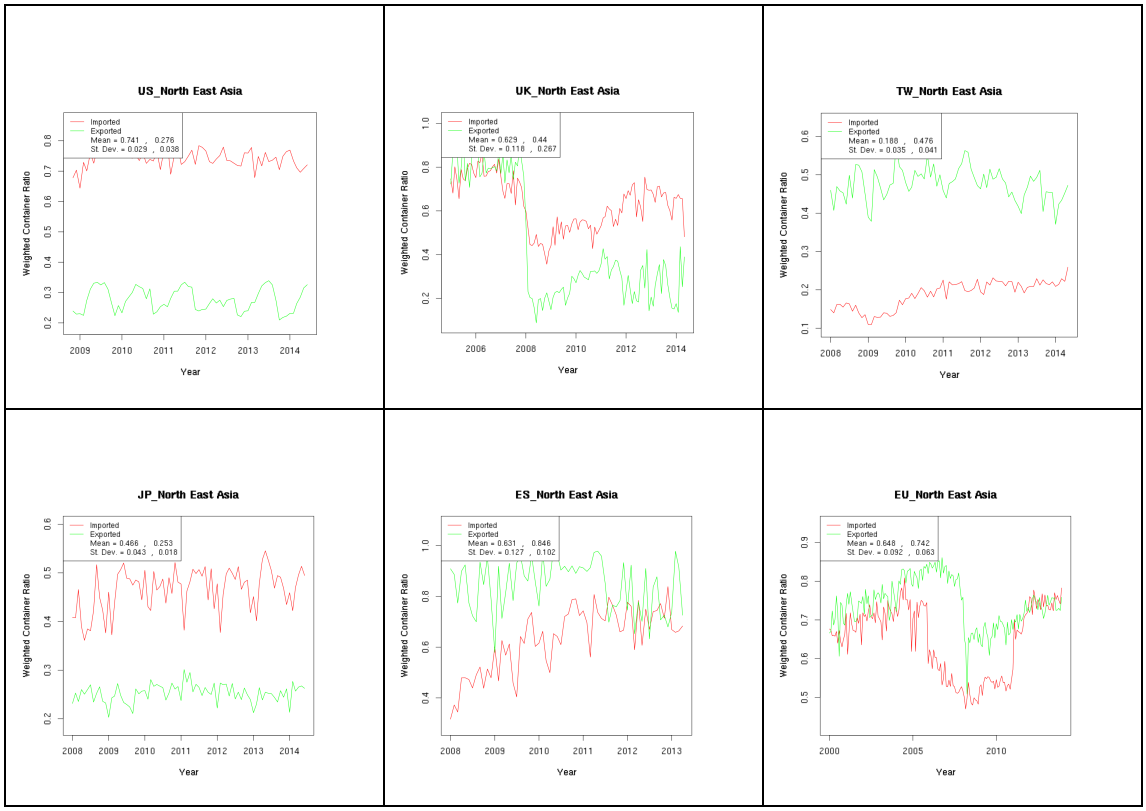


Figure B.6: The importing and exporting CR's of the different datasets to North East Asia.

Finally, the region North East Asia is considered, illustrated in Figure B.6. Again the trade volumes are relatively large and its location is next to South East Asia. Although their geographic location are next to each other, the similarities between them are small. Furthermore, there seems to be seasonality present among all the datasets. This has to do with the Chinese New Year and the type of commodities shipped from the region. The drop in 2008 is both present in the United Kingdom and the Eurostat data. Moreover, trends among the series are small. Only in Spain trends can be spotted, but due to small trade volumes it could be questioned whether this behaviour is structural.

B.4 Correlation weight and volume for chemicals group

The methodology of Section 4.6 was also performed on the chemicals group. The G2 group "FF" represents commodities, which do not belong to other G2 classes. It is a rather big group described as "Other chemicals". It is expected that there also might be a correlation between them, although the relationship would be not as significant.

Country	Estimated	Standard	t-value	P(>|t|)	R-squared

	Coefficient	Error			
US Export **US Import**	2.791e-08 1.334e-08	1.094e-08 2.693e-08	2.552 0.495	0.0131 * 0.622	0.09105 0.00376
UK Export **UK Import**	1.745e-06 -3.815e-08	7.274e-07 2.640e-07	2.399 -0.144	0.0181 * 0.885	0.04928 0.000188
ES Export **ES Import**	-2.584e-07 -1.806e-07	5.408e-08 3.329e-08	-4.778 -5.424	1.13e-05 *** 1.02e-06 ***	0.2691 0.3218
TW Export **TW Import**	-1.200e-07 -4.520e-08	3.258e-08 2.407e-08	-3.683 -1.878	0.00043 *** 0.0643	0.1532 0.04491
JP Export **JP Import**	-1.044e-07 -6.846e-08	1.487e-08 3.185e-08	-7.024 -2.15	7.9e-10 *** 0.0348 *	0.3937 0.05732
EU Export **EU Import**	-5.653e-10 5.037e-12	2.070e-10 2.638e-10	-2.731 0.019	0.00699 ** 0.985	0.043 0.000

Table B.1: This tables denotes the results of regressing the average shipped weight in the FF category against the weighted CR. Significance codes: 0 "*" 0.001 "**" 0.01 "*" 0.05 "." 0.1 " "**

Table B.1 denotes the results of the regressions within the FF group. The correlation found earlier in the GA group, denoted in Table 5, cannot be observed in this category. Only the import and export of Spain and export of both Japan and Taiwan show. Spain in general has small trade flows, especially in this category. Hence, some of outliers could have an effect on this results. However, the relationship is not as clear as in the dry bulk group.

Again, the significant relationships are plotted in Figure B.7. The low R-squared immediately strike the eye, in the form of the large range of observations around the regression lines. However, the residuals seem to be evenly spread and the downward trend is defendable. Moreover, the significant numbers do also meet the technical conditions.

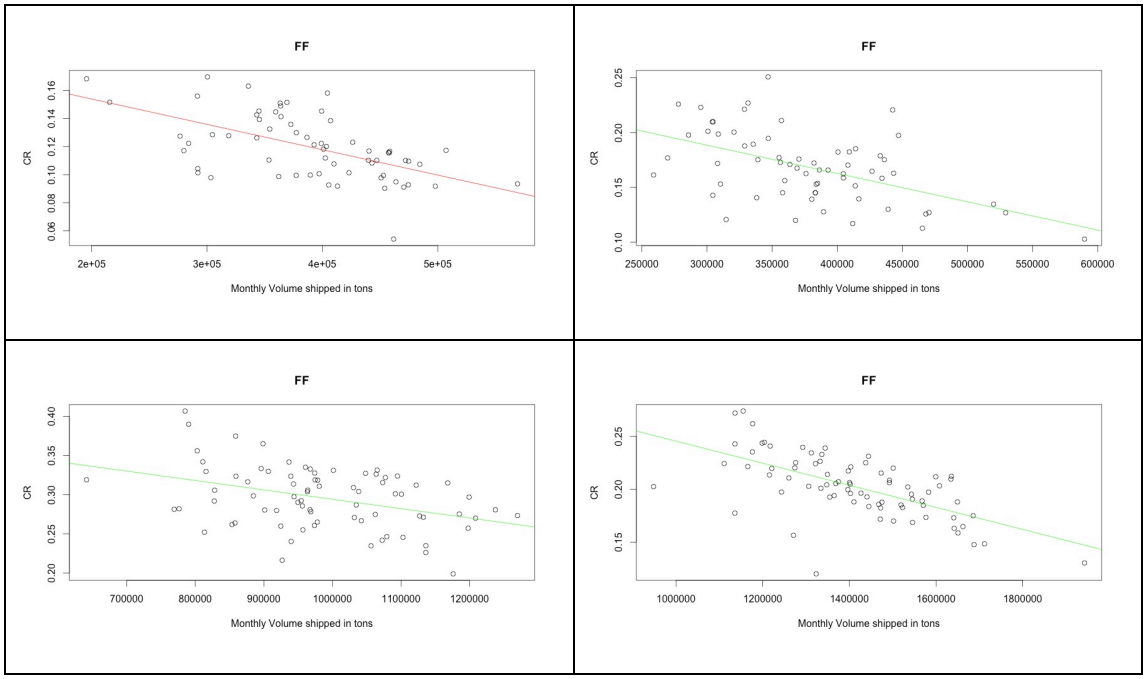
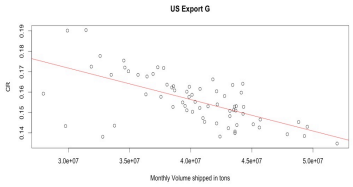


Figure B.7: Selected significant correlations between the CR and weight for the "FF" group. In the top left corner the imported goods in ES, in the top-right corner the exported goods from ES, in the bottom left corner TW exported and bottom right JP export goods.

B.5 Details of the regressions of Section 4.6

G Export US



Call:

lm(formula = CR_Y_E ~ CR_X_E)

Residuals:

Min	1Q	Median	3Q	Max
-0.0293469	-0.0044863	0.0009963	0.0059157	0.0208698

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.180e-01	9.482e-03	22.989	< 2e-16 ***
CR_X_E	-1.542e-09	2.340e-10	-6.591	9.1e-09 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

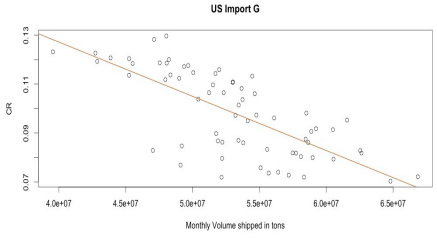
Residual standard error: 0.009479 on 65 degrees of freedom

Multiple R-squared: 0.4006, Adjusted R-squared: 0.3914

F-statistic: 43.44 on 1 and 65 DF, p-value: 9.098e-09

•

G Import US



Call:

lm(formula = CR_Y_I ~ CR_X_I)

Residuals:

Min	1Q	Median	3Q	Max
-0.030092	-0.006570	0.002922	0.008766	0.020316

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.159e-01	1.431e-02	15.090	< 2e-16 ***
CR_X_I	-2.218e-09	2.679e-10	-8.278	9.35e-12 ***

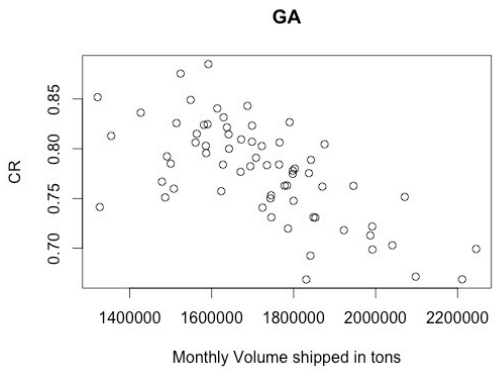
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.01208 on 65 degrees of freedom

Multiple R-squared: 0.5132, Adjusted R-squared: 0.5057

F-statistic: 68.53 on 1 and 65 DF, p-value: 9.354e-12

GA import US



Call:

lm(formula = CR_Y_I ~ CR_X_I)

Residuals:

Min	1Q	Median	3Q	Max

-0.102235 -0.024195 0.007367 0.024032 0.086246

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 1.068e+00 4.054e-02 26.335 < 2e-16 ***

CR_X_I -1.689e-07 2.340e-08 -7.217 7.17e-10 ***

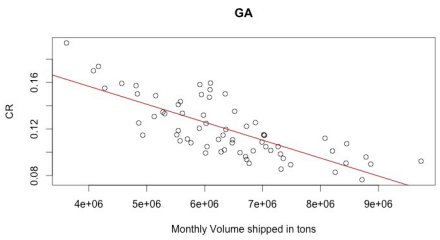
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.0369 on 65 degrees of freedom

Multiple R-squared: 0.4448, Adjusted R-squared: 0.4363

F-statistic: 52.08 on 1 and 65 DF, p-value: 7.169e-10

GA Export US



Call:

lm(formula = CR_Y_E ~ CR_X_E)

Residuals:

Min 1Q Median 3Q Max

-0.027613 -0.012581 -0.001287 0.011279 0.035244

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 2.183e-01 1.054e-02 20.703 < 2e-16 ***

CR_X_E -1.543e-08 1.624e-09 -9.502 6.54e-14 ***

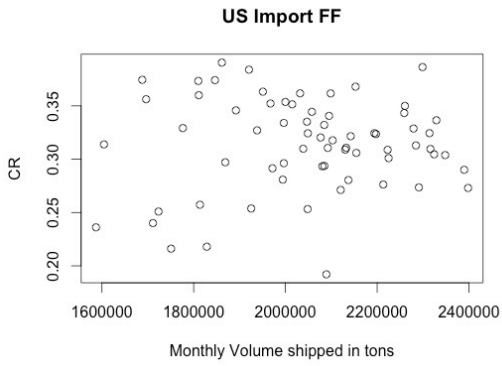
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.01639 on 65 degrees of freedom

Multiple R-squared: 0.5814, Adjusted R-squared: 0.575

F-statistic: 90.28 on 1 and 65 DF, p-value: 6.538e-14

FF Import US



Call:

lm(formula = CR_Y_I ~ CR_X_I)

Residuals:

Min 1Q Median 3Q Max

-0.122678 -0.021446 0.005598 0.031552 0.078853

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 2.869e-01 5.539e-02 5.179 2.34e-06 ***

CR_X_I 1.334e-08 2.693e-08 0.495 0.622

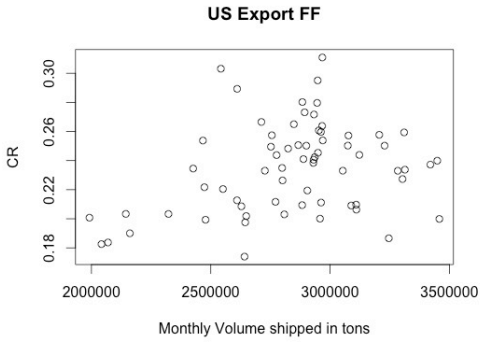
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.04363 on 65 degrees of freedom

Multiple R-squared: 0.00376, Adjusted R-squared: -0.01157

F-statistic: 0.2453 on 1 and 65 DF, p-value: 0.6221

US Export FF



Call:

lm(formula = CR_Y_E ~ CR_X_E)

Residuals:

Min	1Q	Median	3Q	Max
-0.059792	-0.021083	0.000963	0.015379	0.076399

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	1.558e-01	3.129e-02	4.979	4.98e-06	***
CR_X_E	2.791e-08	1.094e-08	2.552	0.0131	*

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

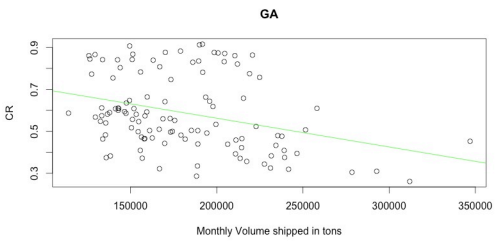
Residual standard error: 0.02948 on 65 degrees of freedom

Multiple R-squared: 0.09105, Adjusted R-squared: 0.07707

F-statistic: 6.511 on 1 and 65 DF, p-value: 0.01308

GA Group UK

Export



Residuals:

Min	1Q	Median	3Q	Max
-0.29201	-0.12028	-0.04533	0.14799	0.34169

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	8.368e-01	7.032e-02	11.901	< 2e-16	***
CR_X_E	-1.372e-06	3.778e-07	-3.632	0.000427	***

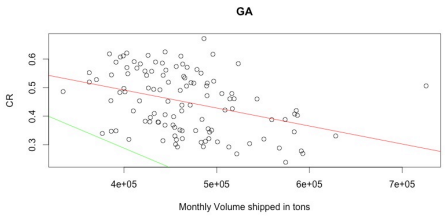
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1693 on 111 degrees of freedom

Multiple R-squared: 0.1062, Adjusted R-squared: 0.09819

F-statistic: 13.19 on 1 and 111 DF, p-value: 0.0004266

GA UK Import



Call:

lm(formula = CR_Y_I ~ CR_X_I)

Residuals:

Min	1Q	Median	3Q	Max
-0.036929	-0.011174	-0.002694	0.011881	0.044117

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	2.535e-01	2.796e-02	9.067	9.94e-14	***
CR_X_I	-1.971e-08	1.027e-08	-1.919	0.0588	.

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.01729 on 76 degrees of freedom

Multiple R-squared: 0.0462, Adjusted R-squared: 0.03365

F-statistic: 3.681 on 1 and 76 DF, p-value: 0.05878

GA group ES Import

Call:

lm(formula = CR_Y_I ~ CR_X_I)

Residuals:

Min	1Q	Median	3Q	Max
-0.032635	-0.009799	-0.001160	0.008143	0.043989

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	2.350e-01	8.155e-03	28.82	<2e-16	***
CR_X_I	-1.185e-07	8.664e-09	-13.68	<2e-16	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.01469 on 62 degrees of freedom

Multiple R-squared: 0.7512, Adjusted R-squared: 0.7472

F-statistic: 187.2 on 1 and 62 DF, p-value: < 2.2e-16

GA EU Import

Call:

lm(formula = CR_Y_I ~ CR_X_I)

Residuals:

Min	1Q	Median	3Q	Max
-0.138869	-0.056710	0.000932	0.054300	0.129569

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	4.721e-01	1.608e-02	29.364	< 2e-16	***
CR_X_I	-1.528e-09	3.972e-10	-3.847	0.00017	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.06927 on 166 degrees of freedom

Multiple R-squared: 0.08187, Adjusted R-squared: 0.07634

F-statistic: 14.8 on 1 and 166 DF, p-value: 0.00017

GA group Exp

EU Call:

lm(formula = CR_Y_E ~ CR_X_E)

Residuals:

Min	1Q	Median	3Q	Max
-0.196580	-0.059319	0.006857	0.058002	0.162367

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	6.227e-01	1.794e-02	34.702	<2e-16	***
CR_X_E	2.281e-10	5.734e-10	0.398	0.691	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.08191 on 166 degrees of freedom

Multiple R-squared: 0.0009519, Adjusted R-squared: -0.005066

F-statistic: 0.1582 on 1 and 166 DF, p-value: 0.6914

GA ES Export

Call:

lm(formula = CR_Y_E ~ CR_X_E)

Residuals:

Min	1Q	Median	3Q	Max
-0.17822	-0.05657	-0.01193	0.06536	0.19521

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	6.511e-01	5.040e-02	12.918	<2e-16	***
CR_X_E	-5.703e-07	2.307e-07	-2.472	0.0162	*

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.08401 on 62 degrees of freedom

Multiple R-squared: 0.0897, Adjusted R-squared: 0.07502

F-statistic: 6.109 on 1 and 62 DF, p-value: 0.01621

GA group Import TW

Call:

lm(formula = CR_Y_I ~ CR_X_I)

Residuals:

Min 1Q Median 3Q Max

-0.16243 -0.07543 -0.01388 0.02815 0.31730

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 7.402e-01 6.944e-02 10.660 < 2e-16 ***

CR_X_I -4.544e-07 1.051e-07 -4.321 4.7e-05 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1079 on 75 degrees of freedom

Multiple R-squared: 0.1993, Adjusted R-squared: 0.1887

F-statistic: 18.67 on 1 and 75 DF, p-value: 4.703e-05

GA Export TW

Call:

lm(formula = CR_Y_E ~ CR_X_E)

Residuals:

Min 1Q Median 3Q Max

-0.238576 0.000049 0.004717 0.008498 0.017932

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 9.796e-01 1.438e-02 68.126 <2e-16 ***

CR_X_E -1.631e-07 3.763e-07 -0.434 0.666

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.02931 on 75 degrees of freedom

Multiple R-squared: 0.002499, Adjusted R-squared: -0.0108

F-statistic: 0.1879 on 1 and 75 DF, p-value: 0.665

GA Import JP

Call:

lm(formula = CR_Y_I ~ CR_X_I)

Residuals:

Min 1Q Median 3Q Max

-0.036929 -0.011174 -0.002694 0.011881 0.044117

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 2.535e-01 2.796e-02 9.067 9.94e-14 ***

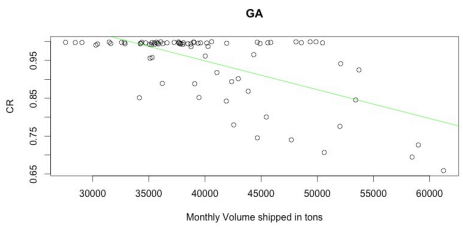
CR_X_I -1.971e-08 1.027e-08 -1.919 0.0588 .

Residual standard error: 0.01729 on 76 degrees of freedom

Multiple R-squared: 0.0462, Adjusted R-squared: 0.03365

F-statistic: 3.681 on 1 and 76 DF, p-value: 0.05878

GA Export JP



Call:

lm(formula = CR_Y_E ~ CR_X_E)

Residuals:

Min 1Q Median 3Q Max

-0.16789 -0.03195 0.01253 0.03886 0.12745

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 1.254e+00 4.495e-02 27.889 < 2e-16 ***

CR_X_E -7.621e-06 1.099e-06 -6.932 1.18e-09 ***

Residual standard error: 0.06959 on 76 degrees of freedom

Multiple R-squared: 0.3873, Adjusted R-squared: 0.3793

F-statistic: 48.05 on 1 and 76 DF, p-value: 1.18e-09

FF GROUP

> summary(res_I_FF_ES)

Call:

lm(formula = CR_Y_I ~ CR_X_I)

Residuals:

Min 1Q Median 3Q Max

-0.052371 -0.012257 0.001629 0.012083 0.041145

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 1.899e-01 1.318e-02 14.410 < 2e-16 ***

CR_X_I -1.806e-07 3.329e-08 -5.424 1.02e-06 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.01901 on 62 degrees of freedom

Multiple R-squared: 0.3218, Adjusted R-squared: 0.3109

F-statistic: 29.42 on 1 and 62 DF, p-value: 1.019e-06

> summary(res_E_FF_ES)

Call:

lm(formula = CR_Y_E ~ CR_X_E)

Residuals:

Min 1Q Median 3Q Max

-0.063993 -0.018990 0.000343 0.017521 0.074427

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 2.659e-01 2.073e-02 12.830 < 2e-16 ***

CR_X_E -2.584e-07 5.408e-08 -4.778 1.13e-05 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.02769 on 62 degrees of freedom

Multiple R-squared: 0.2691, Adjusted R-squared: 0.2573

F-statistic: 22.83 on 1 and 62 DF, p-value: 1.125e-05

UK Import FF

> summary(res_I_FF_UK)

Call:

lm(formula = CR_Y_I ~ CR_X_I)

Residuals:

Min 1Q Median 3Q Max

-0.10461 -0.06412 -0.01236 0.05426 0.20462

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 2.909e-01 5.565e-02 5.227 8.17e-07 ***

CR_X_I -3.815e-08 2.640e-07 -0.144 0.885

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.07419 on 111 degrees of freedom

Multiple R-squared: 0.000188, Adjusted R-squared: -0.008819

F-statistic: 0.02088 on 1 and 111 DF, p-value: 0.8854

UK export FF

> summary(res_E_FF_UK)

Call:

lm(formula = CR_Y_E ~ CR_X_E)

Residuals:

Min	1Q	Median	3Q	Max
-0.24495	-0.11706	-0.05036	0.12412	0.36879

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.520e-01	9.569e-02	1.589	0.1150
CR_X_E	1.745e-06	7.274e-07	2.399	0.0181 *

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1524 on 111 degrees of freedom

Multiple R-squared: 0.04928, Adjusted R-squared: 0.04072

F-statistic: 5.754 on 1 and 111 DF, p-value: 0.01812

summary(res_I_FF_TW)

Call:

lm(formula = CR_Y_I ~ CR_X_I)

Residuals:

Min	1Q	Median	3Q	Max
-0.063214	-0.015756	-0.000158	0.016108	0.066518

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.195e-01	2.217e-02	14.413	<2e-16 ***
CR_X_I	-4.520e-08	2.407e-08	-1.878	0.0643 .

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.02696 on 75 degrees of freedom

Multiple R-squared: 0.04491, Adjusted R-squared: 0.03217

F-statistic: 3.526 on 1 and 75 DF, p-value: 0.06429

> summary(res_E_FF_TW)

Call:

lm(formula = CR_Y_E ~ CR_X_E)

Residuals:

Min	1Q	Median	3Q	Max
-0.086610	-0.024023	0.003425	0.026701	0.086870

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.140e-01	3.236e-02	12.792	< 2e-16 ***
CR_X_E	-1.200e-07	3.258e-08	-3.683	0.000432 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.0361 on 75 degrees of freedom

Multiple R-squared: 0.1532, Adjusted R-squared: 0.1419

F-statistic: 13.57 on 1 and 75 DF, p-value: 0.0004317

JAPAN

Call:

lm(formula = CR_Y_I ~ CR_X_I)

Residuals:

Min	1Q	Median	3Q	Max
-0.098903	-0.020163	0.003366	0.021107	0.063960

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	5.152e-01	2.832e-02	18.19	<2e-16 ***
CR_X_I	-6.846e-08	3.185e-08	-2.15	0.0348 *

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.03003 on 76 degrees of freedom

Multiple R-squared: 0.05732, Adjusted R-squared: 0.04492

F-statistic: 4.621 on 1 and 76 DF, p-value: 0.03476

```
> summary(res_E_FF_JP)

Call:

lm(formula = CR_Y_E ~ CR_X_E)

Residuals:

      Min        1Q    Median        3Q       Max

-0.091694 -0.011446 -0.000176  0.015328
0.044631

Coefficients:

              Estimate Std. Error t value Pr(>|t|)

(Intercept)  3.499e-01  2.117e-02  16.530  <
2e-16 ***

CR_X_E      -1.044e-07  1.487e-08  -7.024
7.9e-10 ***

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.'
0.1 ' ' 1

Residual standard error: 0.02303 on 76
degrees of freedom

Multiple R-squared:  0.3937,
                                    Adjuste
d R-squared:  0.3857

F-statistic: 49.34 on 1 and 76 DF,  p-value:
7.899e-10
```

EU Export FF

```
Call:

lm(formula = CR_Y_E ~ CR_X_E)


Residuals:

      Min        1Q    Median        3Q       Max

-0.129825 -0.047763 -0.001264  0.042960
0.114028

Coefficients:

              Estimate Std. Error t value Pr(>|t|)

(Intercept)  6.161e-01  1.120e-02  55.029  <
2e-16 ***

CR_X_E      -5.653e-10  2.070e-10  -2.731
0.00699 **

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.'
0.1 ' ' 1

Residual standard error: 0.0522 on 166
degrees of freedom

Multiple R-squared:  0.043,         Adjusted
R-squared:  0.03724

F-statistic: 7.459 on 1 and 166 DF,  p-value:
0.006994
```

EU Import FF

```
Call:

lm(formula = CR_Y_I ~ CR_X_I)

Residuals:

      Min        1Q    Median        3Q       Max

-0.09366 -0.04012 -0.01285  0.03255  0.11914

Coefficients:

              Estimate Std. Error t value Pr(>|t|)

(Intercept) 3.705e-01  1.127e-02  32.872
<2e-16 ***

CR_X_I      5.037e-12  2.638e-10   0.019
0.985

---

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.'
0.1 ' ' 1

Residual standard error: 0.05241 on 166
degrees of freedom

Multiple R-squared:  2.197e-06,     Adjusted
R-squared:  -0.006022

F-statistic: 0.0003647 on 1 and 166 DF,  p-
value: 0.9848
```