Data preparation for regional logistics data and the complexities involved

Chua Wan Theng; Jouta Lim Zi Yu; Lin Qian Pin

Singapore Management University

ABSTRACT

With the increasing regional presence of logistics providers, the logistics data used by them have been increasingly complex for the different contexts involved. Unlike a logistics provider providing services for just one country, regional logistics providers will face the complexities involved in having geospatial address data in different formats, and data records of different languages corresponding to the local cultures of the different locales. The lack of a universal uniform format in writing affects the data recorded in the system, which becomes a complication faced when running analysis on such data. In this paper, we will mainly discuss on the complexities in geocoding, for the different address formats in the different countries, and the differences in geospatial data accuracy, as a result of the different countries using different methods in anchoring address fields to physical geographical locations. Data records of different languages, and different vendors employed to fulfill the logistics services also add on to the complexity of the data, for the lack of uniformity necessitates the need to prepare the data in a uniform format to be able to better draw on the insights in the data for analysis. The paper seeks to highlight the complexities involved in regional logistics data and discuss the methods we have applied to overcome them.

INTRODUCTION

The growing influence of the internet in our lives have led to an increase in avenue for eCommerce platforms. Along with the growth of eCommerce, logistics firms whom are involved in the transport and distribution of products have started to play an increasingly important role in the industry. Today, logistics firms go beyond their traditional role of providing services for the B2B market, but have been enhancing their services to better cater to the B2C and C2C markets that have been gaining dominance in recent years. Logistics firms have gained opportunities to expand beyond their shores, to set up services across different regions serving customers across the globe.

This research paper seeks to focus on the complexities in logistics data that comes as a part of being a regional entity in the logistics market, serving different countries with different languages and norms. For our research, we worked together with a regional logistics provider who has sponsored us with their data for our investigation into the real world data collected for the distribution services they have provided. The data given to us includes a number of countries, such as Australia, New Zealand, Japan, Korea, Hong Kong, Taiwan and Thailand. Not all of these countries use English as their main language of communication, some of them use their native language in their daily operations that becomes captured in the data collected. The different countries also differ in their norms pertaining logistics processes, creating a need to make adjustments to fit the different locales as things are not always in a one size fits all situation. In the paper, we will discuss in detail the different complexities faced in the data and how we sought to address the issues faced, and the design of our dashboard for the analysis of such operations data.

LITERATURE REVIEW

To better understand the processes and applications of geocoding data, we have done some literature review on relevant sources to gain more background knowledge and understanding of the subject before we start on the analysis on the data provided.

In Goldberg, Swift, and Wilson (2008), they discussed that most of the geospatial analysis involves polygons and the mapping of data over to these polygons to produce meaningful maps. It is important to be able to represent not only meaningful but accurate and appealing information. Such can be controlled by the size of the polygons. Smaller polygons would definitely produce the most accurate and meaningful information to any user as the granularity of detail provides a thorough understanding of the information. As polygons start to represent larger geographic objects such as States or the entire country, the dataset quickly becomes less accurate.

In Goldberg and Cockburn (2012), they stated that geocoding have been used in numerous countries to conduct geospatial analysis. However, there are numerous limitations and problems to spatial accuracy. Firstly, false clusters can occur whereby same postal codes are used resulting in a cluster existing when in reality, better geocodes could be determined from the street addresses. Secondly, when clustering, a centroid may lie outside of the boundary of the polygon if it has an irregular shape, and may fall within the boundary of a neighbouring polygon. Additionally, there might be issues with geocoding whereby the centroid of an areal unit is placed in a location which might be inapplicable or invalid such as deserts or water bodies. Lastly, there might be postal codes that crosses boundaries when computing county-level analysis, as such it may not be easy as to where to classify the postal code.

In DuVander (2010), the author introduces geocoding as a means to link an address with a physical location, such as the latitude and longitude, to locate an approximate location. The process involves breaking down an address into its different components to use them to identify the unique location associated with the different components. Following this, the author introduces two geocoding methods using either JavaScript or HTML. The author also emphasizes that the addresses used in geocoding would likely be an user input, and provides examples of existing services such as Google and how they make adjustments to the input to better approximate the location. The main takeaway from this ebook would be the breakdown of geocoding accuracy, ranging from the wider classifications such as country, state and county to specifics such as street, intersection and building. Despite the different countries and contexts around the world, this provides a form of a template for one to understand how addresses and locations are coded to a certain accuracy. The author also introduces the concept of reverse geocoding, whereby one seeks to find an address from a given location, a reverse of the usual geocoding process. The example he takes from google shows that if a clear address to denote the exact point cannot be found, a best approximate would be taken instead. This highlights the key difference between addresses and geographical locations, for that geographical locations can be narrowed to a single coordinate based on its longitude and latitude, whereas addresses are based upon landmarks created by people to denote a specific location. Geocoding is not a precise science, for that addresses refer to points that make sense, but postal codes aim to achieve some accuracy in using the center point of a location as a reference in its coding.

COMPLEXITIES IN GEOCODING BY COUNTRY

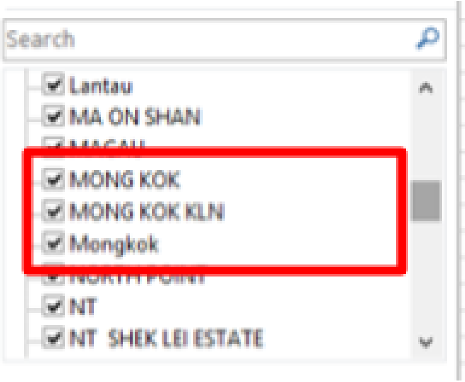
The shipment data provided to us includes geospatial data pertaining to the sender and receiver addresses. These properties made it appropriate for us to geocode the data in order to perform detailed analysis to improve processes. However, before such data could be utilised effectively, the intermediate steps of cleaning and comprehending has to be performed.

1.1 HONG KONG

A major restriction in conducting geospatial analysis for Hong Kong’s data is the lack of a proper national postal code system in Hong Kong. Due to the inability to easily geocode a shipment, attempts to geocode a shipment using information such as address and district have also proved futile.

This is due to the inconsistency in data recording across employees and customers.

For example in figure 1, there are 3 different records of Mong Kong, a district in Hong Kong. Also, there are also differences in the way addresses are recorded. Addresses could be seen recorded in full with the unit and block number, whereas there are other instances where addresses only contained of the street name or building name.



Search

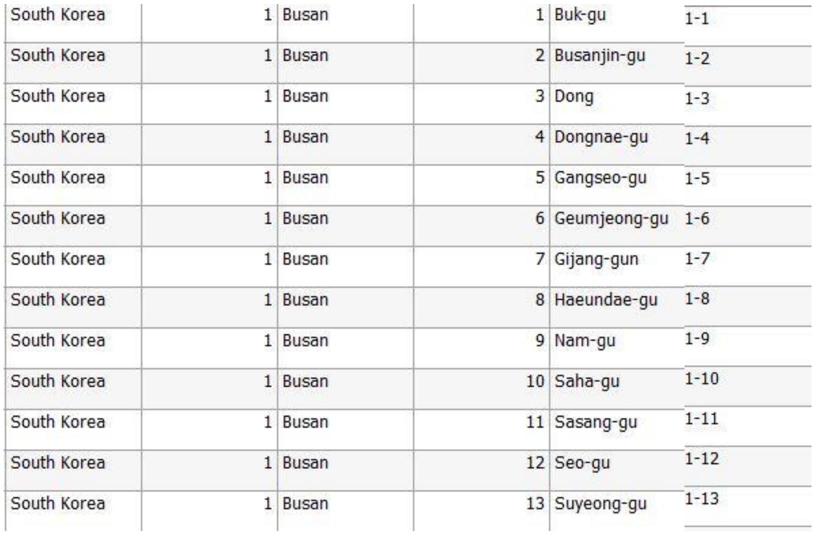
☑ Lantau
☑ MA ON SHAN
☑ MACAU
☑ MONG KOK
☑ MONG KOK KLN
☑ Mongkok
☑ NORTH POINT
☑ NT
☑ NT SHEK LEI ESTATE

Figure 1: District Names in Hong Kong

1.2 KOREA

Problems arose in both the sponsor's and publicly available data. Korea used to operate on a 6-digit postal code system, but as of August 1st, 2015, the country switched to a new 5-digit postal code system. However, the sponsor's data had both the old and new system present in their data, and this unnecessarily complicated the geocoding process.

A workaround this problem was to make use of the state and city names. The sponsor's data had information on Receiver's and Sender's City, State, and Address. Fortunately, all State names were unique. (E.g. Busan, Incheon). Tying the unique State name together with the City residing within, we were able to generate our own unique national ID for each City in Korea. The naming of States and Cities were consistent across both the sponsor's dataset as well as the publicly available data. This approach was adopted to bypass the complexities of two different postal code system residing in the sponsor's dataset.



South Korea	1	Busan	1	Buk-gu	1-1
South Korea	1	Busan	2	Busanjin-gu	1-2
South Korea	1	Busan	3	Dong	1-3
South Korea	1	Busan	4	Dongnae-gu	1-4
South Korea	1	Busan	5	Gangseo-gu	1-5
South Korea	1	Busan	6	Geumjeong-gu	1-6
South Korea	1	Busan	7	Gijang-gun	1-7
South Korea	1	Busan	8	Haeundae-gu	1-8
South Korea	1	Busan	9	Nam-gu	1-9
South Korea	1	Busan	10	Saha-gu	1-10
South Korea	1	Busan	11	Sasang-gu	1-11
South Korea	1	Busan	12	Seo-gu	1-12
South Korea	1	Busan	13	Suyeong-gu	1-13

Figure 2: Geocodes to create Korea's Kcodes

Lastly there also a related issue pertaining to the geospatial data accuracy of the publicly available shape file. The latest version was the 2015 version and there were issues with the polygons. A single entity in the public dataset was linked to 2 polygons erroneously. More specifically, selecting Jung-gu in Busan also led to Jung-gu in Incheon being selected. (as demonstrated in the figure 3 below). This issue did not emerge at the first few steps of cleaning, it was only after the mapping over of data resulted in always the same city always being left out (Busan's Jung-gu) were we made aware of it.

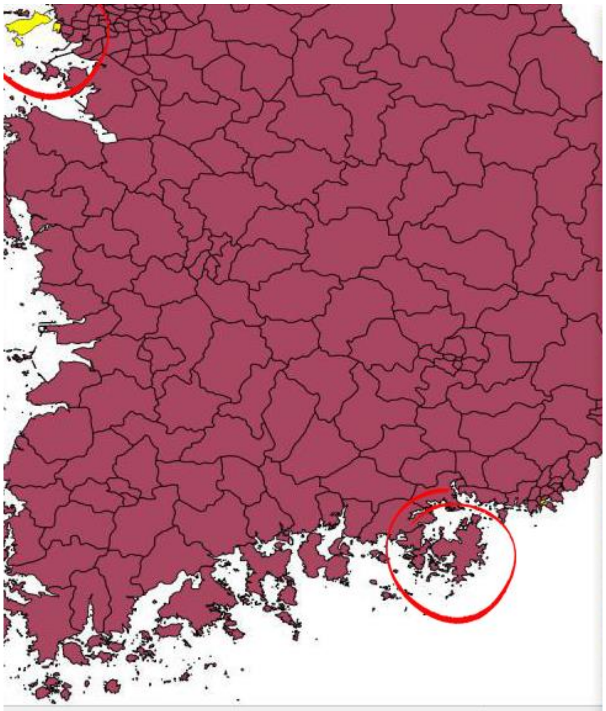


Figure 3: Jung-gu's in Incheon and Busan

1.3 JAPAN

In contrast to the above countries, Japan has implemented a systematic standard in the format for addresses along with its postal code systems, which made it easy to geocode the addresses in the data given to us.

The Japanese address follows a clear format of the following example:

〒123 - 4567 東京都 新宿区 西早稲田 X 丁目 YY-ZZ	〒123 - 4567 X-YY-ZZ Nishiwaseda Shinjuku-ku, Tokyo-to

In the writing of the address, they make clear distinctions between the different levels of classification. This clear distinction is reflected in the forms for documentation used by their national offices which allow you to fill in the addresses in the systematic order from the 都道府県 (prefecture), to 市区町村 (city/ward/town/county) then to the street number and room number if available.

The first line of the address in both languages refers to the postal code.

The second line of the address in Japanese refers to the prefecture, which uses the suffixes of 都道府県 with the name of the prefecture. For example, Tokyo would be written as Tokyo-to, whereas Osaka would be written as Osaka-fu.

The third line of the address in Japanese refers to the city/ward/town/county, which uses the suffixes of 市区町村 with the name. For example, the 23 special wards of Tokyo follow with the -ku suffix, such as Shinjuku-ku, Taito-ku.

The fourth line of the address in Japanese refers to the street level, which is similar to how we write our addresses in Singapore as 81 Victoria Street or Ang Mo Kio Avenue 3. The X 丁目 refers to Avenue X, whereas the YY-ZZ refers to the specific lot number to locate the exact location of the address, similar to the 81 of 81 Victoria Street to find the exact lot on the street.

The systematic writing of the addresses in Japan with clear distinction over the different geographic classification layers made it easy for clear identification of the actual location that correspond to the address.

Furthermore, the postal code is accurate to the avenue name, allowing for one to narrow down to a specific locality such as 西早稲田 X 丁目, which would narrow down to the local post offices in charge of the parcel to find out the exact address of the receiver should the postal code and some ambiguous writings be the only address written for a postal item.

The postal codes themselves also follow a fixed format in the identification of the prefecture level and city level names, using the first 3 to 5 digits of the postal code. This postal code data is published by the Japan Post Office online, allowing for one to download the data to use for geocoding purposes. By using the postal code data, the prefecture and city names were effectively identified from the postal code provided in the sponsor’s data. The data published by the Japan Post Office is regularly updated and maintained, allowing all data to be kept up to date with changes made.

COMPLEXITIES IN DATA PROVIDED

Due to the differences in reporting nature in the data across different countries as well as the complexity of multi-national logistic operations, multiple problems were faced in the data provided by our sponsors.

1.1 DIFFERENT DATA SYSTEMS UTILIZED

Two different logistics tracking systems, App1 and App2 have been utilised by our sponsors in managing logistic shipments in different countries. App1 and App2 are the systems that record the data pertaining to the shipment details such as the sender and receiver information, and tracks the distribution process for each shipment from the picking up of the shipment to its successful delivery. Japan, Korea, Taiwan, Thailand and Hong Kong uses App1, whereas Australia and New Zealand uses App2. There exists a number of differences in the reporting of data across these two apps.

As the data structure across both apps are different, this makes the reconciliation of data across both apps difficult and impossible. This is due to the lack of consistency in naming different columns, the lack of specific columns of information in App2 which makes understanding the problems of a shipment more difficult compared to App1.

1.2 DIFFERENCES ACROSS THE DIFFERENT COUNTRIES

As with differences in cultures in different countries, this thus resulted in differences in recording data across different countries. Below, we will explain in detail how the handling of these regional data adds new layers of complexity beyond what we would have had experienced in the classroom with the data given to us by our school.

1.2.1 DIFFERENT LANGUAGES USED

In Singapore, we see that we have standardized all official processes to be done and documented in the English language for that it is the official language of our nation. Similarly, other countries would choose to document their work in their official language, which may not always be English. This adds another layer of complexity in our data, for that not all countries follow with inputting their data in English. Although system interfaces have been designed in English with some fields to be picked from a list in English, some information such as the addresses and names which were originally written in another language by the end consumer has to be inputted in exactly into the system.

For the list of countries that we have the data for, the data collected includes that of English, Japanese, Chinese, Thai and Korean languages. Although not all fields written in non-English languages may be key figures used in determining the key performance indicators, some of them involve data that would be used in further analysis. Main data fields that were written in different languages were those pertaining to geospatial data, for that addresses could be written in local languages instead of English. This affected us in using some of the data fields originally in the data. For example, in the data for Japan, the column that records the state of an address has that of both languages of Tokyo and 東京都, and some cases it was written in different formats such as Tokyo and Tokyo-to. The variations across languages and the corresponding variations when written in English saw a need to be reconciled in a singular standard format to be derived from the data we have.

For countries with a postal code system and their postal code data freely available online and is up-to-date, we were able to derive a standardized list of values in English using these sources and matched it to our data columns which consists of postal code values. Postal code values were written in numerical formats, free from the languages, hence allowing us to use as an accurate reference in deriving the required data columns of state and city level details in English from external sources.

The difference in languages not only affects our ability to assess the data directly, but also in accessing data sources available online that may be specific to their national tongue. Used to the westernized society of Singapore, we may be inclined to think that all official data on the internet such as our own data.gov.sg will necessarily be in English. However, that had not been the case when we had to source for data for our usage. For example, the Japan Post Office had posted their postal code data online but could only be accessed on their Japanese website. On the other hand, Korea’s geospatial data files were on their equivalent of our data.gov.sg, however was in Korean and required identifications as a Korean resident in order to gain access to it. This challenge faced is not just limited to us who are working on this project, but also applies to the sponsors also had to work with local entities or speakers of the local languages to gain access to data in their implementation of the different projects. Despite the challenges faced in this, it also allowed us to gain exposure to such data that we would not have the opportunity to unless we were working on regional data, which proved useful to us in preparing ourselves and also learning how to handle it despite our lack of knowledge in the languages.

1.2.2 DIFFERENT WAYS OF IDENTIFYING SHIPMENT TYPE AS INBOUND OR OUTBOUND

Through our experience in the project, we understood that there are different ways in identifying an inbound or outbound shipment for different countries and for different apps.

The major difference lies in countries using App1. Firstly, for countries using App1. In Japan, a specific string (e.g. ABC XXX) in the ‘Receiver Name’ is used to identify inbound shipments. If the name in ‘Receiver Name’ matches our intended string, it is labelled as an inbound shipment, else outbound. This method is similarly used in Korea and Taiwan, however the difference would be the specific string used to identify. On the other hand, Thailand identifies its outbound shipments using the “Sender Address”. As such, if the sender address matches the address of the company address in Thailand, it is thus an outbound shipment. Lastly, Hong Kong identifies its inbound shipments using the “Customer Account No”. If the customer account number matches the specific account number for our sponsor, it is thus labelled as an inbound shipment. Other account numbers are labelled as outbound shipments.

On the other hand, App2 which is used by Australia and New Zealand identifies their inbound and outbound shipments using the “Pickup Address” and “Receiver Address”. If the “Pickup Address” matches the warehouse’s address, it is an outbound shipment. Whereas if the “Receiver Address” matches the warehouse’s address, it is then an inbound shipment.

As such, the various differences in simply identifying an inbound or outbound shipment for further analysis suggests the complexities managing the data. This also proves the inflexibility in creating a unified dashboard for all different countries.

1.3 DIFFERENT VENDORS/SERVICE PROVIDERS

Different vendors and service providers may be used for different shipments and also in the different countries. Some vendors specialize in handling bulk shipments, while others may differ in the areas they service, hence our sponsors engage a number of different vendors in their logistics services, and integrates the shipment data from the vendors into their own systems of App1 or App2. The engagement of different vendors and the assessment of their performances may differ, such that there might be different service level agreements (SLA) made between the different vendors despite the same areas handled by them. For example, Vendor A might be able to deliver to State XYZ within 1 day, however Vendor B might only be able to do so within 2 days, hence they cannot be assessed on one singular standard but has to be assessed based on what they have agreed as per the SLA with our Sponsor. Next, we will explain in detail the complexities resulting from the engagement of multiple vendors.

The integration process of shipment data between the vendors and our sponsor requires extensive status mapping to be done, to find the best match between the statuses provided by the vendor and those currently existing in App1 or App2. The statuses currently in the system may not always be a one-size-fits-all, hence may require the mapping and creation of additional statuses to fit the level of detail in the statuses provided. It is also important to note that the different levels of detail in the statuses of the different vendors may also correspond to the expected level of detail of each locale, which may not be like the Singaporean context which our sponsor’s App1 and App2 were based on. This level of detail may also be affected by the differences in the shipment distribution process by the vendors in response to the needs of the local community, hence needs to be taken into account in assessing the performance of each vendor. Firms handling logistics processes of different regions will have to understand the vendors it engages and the local context in order to better handle the expectations of their clients.

The different vendors engaged may follow different SLA agreed upon between the individual vendors and the sponsors, hence this results in the need to assess the different vendors by the different standards. The Sponsor’s App1 data contains a data column to record “Agent / Vendor name”, which is required in order to make the distinction from the list of shipments on which are handled by which vendor. By matching with the vendor and the corresponding geolocation specified in the SLA, we could compute the key performance indicators following the standards from our sponsor in measuring turnaround time taken to assess the vendors’ performance.

1.4 DIFFERENT DEFINITIONS FOR THE COMPLETION OF A SHIPMENT

As end consumers and users of the services available on the market, we tend to have the perception that the end of the shipment cycle would be when the item is delivered successfully into the hands of the recipient. This may not be an inaccurate definition of the end point, but is only one of the many different possibilities. Below is a flowchart to show an example of possible endpoints that would mark the completion of the shipment.

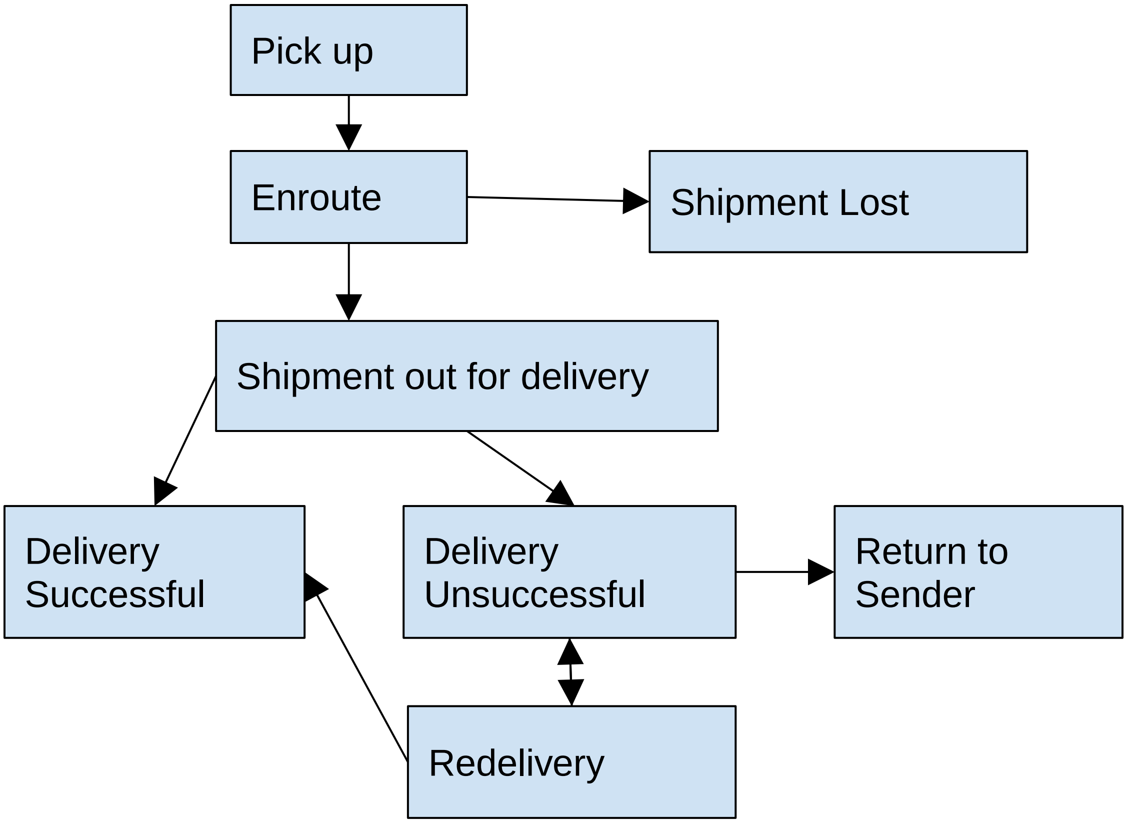


Figure 4: Complete Shipment Process

As we can see from the flowchart, delivery successful may not always be the end point. The shipment could be lost in transporting, which would mark it as an exception case to be investigated. Multiple delivery unsuccessful attempts could mark the item to be returned to the sender, following the standard protocols set out by each logistics service provider in assessing how long will they hold a shipment following unsuccessful delivery attempts.

The representation of these different stages of the shipment's cycle are recorded in the systems as status codes corresponding to each unique status code. The details of what is involved in causing the different statuses are also captured in the specific status codes, such as the number of times a shipment was out for a delivery attempt, exception cases involving shipment damage and the extent of damage. The status codes hence contains these specific details in them to be used in understanding the issue beyond the general level of details shown to the general consumer.

Following the different statuses to be taken into account as potential endpoints of the shipment cycle, we cannot analyse the data in isolation with a single fixed endpoint, but have to understand which statuses are to be accounted for in the data. This requires us to allow for variable change following ways to mark a completion and provide for the flexibility in adjusting for the definitions of the endpoints of the shipment cycle. Perhaps enhancements can be made for existing systems to allow for certain fixed end points, however the dynamics in the data coverage of the different contexts may be eroded should that be done instead.

1.5 DIFFERENT DATA SOURCES REQUIRED TO ASSESS PERFORMANCE

Adding to the complexity of managing data across different systems, our team have also identified a difficulty due to the different data sources required to assess the performance of a shipment. Due to the lack of a consolidated data system in the company, the procedures to prepare the data to assess a shipment's performance is tedious and complicated.

The following is a list of sources required to be consolidated before further analysis could be done.

No	Name of Data Source	Details of Data Source
1	Shipment Info	Details of each shipment. Information includes tracking number, origin and destination, sender and receiver details
2	Status Info	Records of each shipment at each stage of the process. This includes time stamps at different stage codes
3	Stage Codes	Database of all the different stage codes and the definition behind each stage code. There are 6 different main shipping processes - Pickup, Processing, Linehaul, Exceptions, Delivery and Returns. Under each shipping process contains a list of stage codes that represents a unique situation encountered during a shipping process.
4	Service Level Agreement (SLA) Mapping	Database of all the SLA's required for each postal code in each country. This is to identify shipments which have failed whereby their turnaround time have exceeded the expected SLA time.
5	Reason Codes for Failure in First Delivery Attempt	There exists a separate list of reason codes and its description for deliveries which have failed their first attempt. These reason codes are mapped to the specific stage codes in the origin data file.
6	Postal codes Database	This database contains the name of the city, district, latitude and longitude for each country. The geospatial information provided in the original dataset consists only the postal codes. As such, a separate dataset is required to easily identify the location of these postal codes and easily group these postal codes into specific districts.

1.6 UNIQUE GEOSPATIAL MAPPING FOR AUSTRALIA AND JAPAN

Not all countries have progressed on the same developmental stages in mapping geospatial data for their country, some may have done extensive data collection to gather all details of the geographical landscape whereas others may only have basic data regarding the roads only. In addition, the geospatial data collected may not be geocoded in similar manners, dependent on the context and culture of the locality on what attributes have been assigned to such data records. For the purposes of the project, we have focused on the geospatial data of Australia and Japan, the two main territories handled by our sponsor. Due to the factors explained above, the format and attributes of the geospatial data available from these two countries differed greatly.

The geospatial attributes of the data for Japan captured in App1 were the postal codes of the receiver. However, the publicly available shape file for Japan that we sourced for did not use postal codes to identify the polygons denoting the different areas in Japan. Instead, the shape file contained data of the JIS code, to be explained in detail in the later section. By utilizing a separate data source to map the JIS codes to the postal codes, we were able to join the shape file to the shipment data from our sponsor's data and hence map out the geospatial data based on our sponsor's requirements.

While the geospatial mapping was rather straightforward for Japan for they had datasets that were easily convertible to map with each other, it was not the same for Australia's geospatial data. Unlike Japan that had clearly defined areas and methods of classifying these territories, Australia did not have a fixed list of the classifications and definitions of the areas, whereby some definitions were often subjected to arbitrary definitions by the locals. Despite the clear definitions of what consisted as a state in Australia, the suburbs that were referenced in addresses from our sponsor's data were not the nationally recognized standard, hence lacked a complete and up-to-date list for our geospatial mapping. The suburbs used in the sponsor’s data were a mix of either broadly recognized shires and countries or specific neighbourhoods.

In comparison to that, the publicly available shape file that we had sourced used the national standards of Suburbs to name each polygon, which we found to be more standardised and uniform compared to the arbitrary naming standards we saw in the sponsor’s data. Hence to be able to better visualize the sponsor’s data and allow it to conform to widely accepted standards, we have decided to use a separate source to map the latitude and longitude information to the postal codes captured in our sponsor’s data in order to get each individual coordinate mapped on the shape file and then using the shape file’s existing polygon to act as the classifier. By doing so, we were then able to map the shipments to their respective suburbs by the national standards in the map visualization utilized.

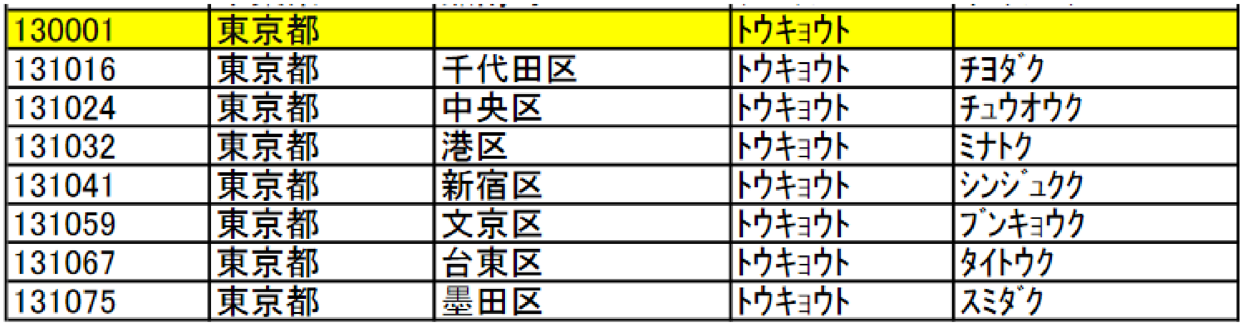
METHODOLOGY

1.1 RESTRICTIONS IN HONG KONG’S GEOSPATIAL ANALYSIS

Due to the limitations in the postal code system in Hong Kong as well as the inconsistencies in recording addresses in the database, it is thus difficult for our team to conduct geospatial analysis until further data resolutions are conducted in the future.

1.2 ALTERNATIVE WAYS TO GEOCODING - JAPAN

Other than the postal code system in Japan, they also have a geocoding system known as the JIS Code (市区町村コード). Unlike postal codes that may see updates following changes in addresses and how postal codes may be assigned separately for the commercial entities beyond geolocation specifications, JIS codes are bound to the addresses by geolocation. JIS codes are also used as identifiers in geospatial data files for Japan unlike postal codes, allowing for greater convenience and compatibility in using them as an identifier.



130001	東京都		ﾄｳｷｮｳﾄ	
131016	東京都	千代田区	ﾄｳｷｮｳﾄ	ﾁﾖﾀﾞｸ
131024	東京都	中央区	ﾄｳｷｮｳﾄ	ﾁｭｳｵｳｸ
131032	東京都	港区	ﾄｳｷｮｳﾄ	ﾐﾅﾄｸ
131041	東京都	新宿区	ﾄｳｷｮｳﾄ	ｼﾝｼﾞｭｸｸ
131059	東京都	文京区	ﾄｳｷｮｳﾄ	ﾌﾞﾝｷｮｳｸ
131067	東京都	台東区	ﾄｳｷｮｳﾄ	ﾀｲﾄｳｸ
131075	東京都	墨田区	ﾄｳｷｮｳﾄ	ｽﾐﾀﾞｸ

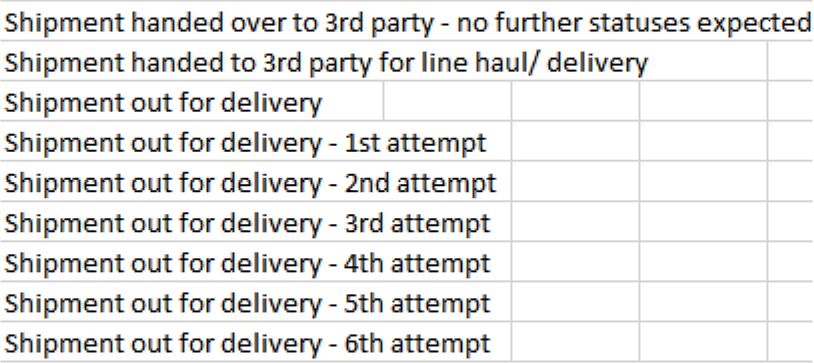
Figure 5: Sample of JIS Codes (from http://www.soumu.go.jp/denshijiti/code.html)

The JIS code system is handled by Japan’s Ministry of Internal Affairs and Communications. The JIS code system assigns a unique number to identify a specific geolocation based on geographical classifications used in the country. For example, JIS code 131041 is Shinjuku ward of Tokyo. This makes it more compatible with geospatial data files such as shapefiles that tends to have polygons on the same level of detail. Hence we have used the JIS code for our geospatial analysis.

1.3 DETERMINING DIFFERENT END POINTS IN A SHIPMENT

One major difference between App1 and App2 is the list of endpoints recorded for each system. App2 provides an advantage is having specific starting and ending points to a shipment. This thus allows the ease in calculating the turnaround time of a shipment in App2. However, with specific starting and ending points, this results in less flexibility in further understanding the process of a shipment. As such, this is incorporated into App1, which contains a list of stage codes in categorising an ending point

However with flexibility, there exists complexity is easily identifying an endpoint or the first delivery attempt of a shipment and thus difficulty in calculating the turnaround time for a shipment as seen in figure 6. As such, our team have come up with the solution in providing flexibility for our sponsors in determining the endpoints for shipments from a check box provided in the dashboard. The dashboard takes into account the endpoints checked and calculates the turnaround time for each shipment from the start till the end.



Shipment handed over to 3rd party - no further statuses expected
Shipment handed to 3rd party for line haul/ delivery
Shipment out for delivery
Shipment out for delivery - 1st attempt
Shipment out for delivery - 2nd attempt
Shipment out for delivery - 3rd attempt
Shipment out for delivery - 4th attempt
Shipment out for delivery - 5th attempt
Shipment out for delivery - 6th attempt

Figure 6: Sample of Delivery Statuses

1.4 MANIPULATING DIFFERENT DATA SYSTEMS

As mentioned and elaborated above in the differences in data structures for App1 and App2, we have thus created two separate dashboards to accommodate the differences. However, the reporting of the information and charts used will be same to provide consistency in data understanding for our clients.

DASHBOARD DESIGN

In this section, we will explain our application design of the dashboard we have created to best fit the shipment data given to us by our sponsor. In this project, we have used R Shiny to create a dashboard to integrate the different insights we have gathered.

R Shiny is a web application framework for R, which allows us to turn our analyses done in R into interactive web applications that can be hosted on a server for easy access by our sponsors. Our choice of using R Shiny is because of its ease of use, and flexibility in integrating different types of charts, as well as it being open-sourced and free. Compared to other commercial platforms available, R Shiny would serve to be a more sustainable platform for our sponsors to use for that it is free and that no web development skills are required, making it easier for them to make changes to fit their situation. An interactive application would best fit the needs of the sponsors, for the easy usage with controls fit to their specifications would suit the needs of the sales team.

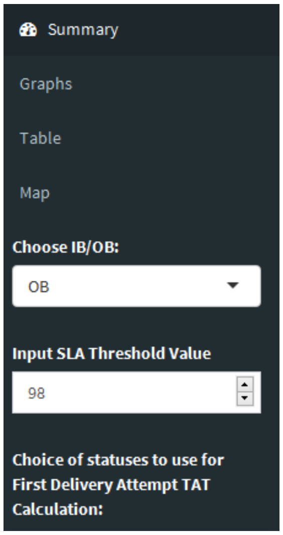
The full dashboard consists of the main body and the sidebar. The sidebar consists of filters and the navigation tabs for the main body. The main body displays the different data visualizations available, such as the graphs, the data tables and the geospatial map. We will explain the

4 main parts to our dashboard below, namely, the Filters, Summary and Graphs, Data table and Geospatial Map.

1.1 FILTERS

To capture the flexibility in determining the start and ending points of a shipment, filters in the form of a check box have been created to allow our sponsors to set the start and end statuses according to their specifications to be taken into the calculation for the turnaround time.

Additionally, an Inbound/Outbound (IB/OB) filter has been created to allow our sponsors to easily filter to those categories of shipments so that they can understand the situation of the shipments respectively.



Choose IB/OB: OB

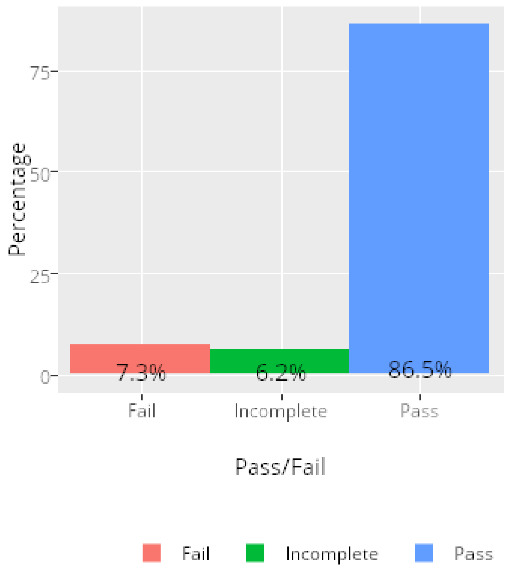
Input SLA Threshold Value: 98

Choice of statuses to use for First Delivery Attempt TAT Calculation:

Figure 7: Screenshot of Dashboard Filters

1.2 GRAPHS

1.2.1 SUMMARY - UNDERSTANDING OF THE OVERALL SITUATION



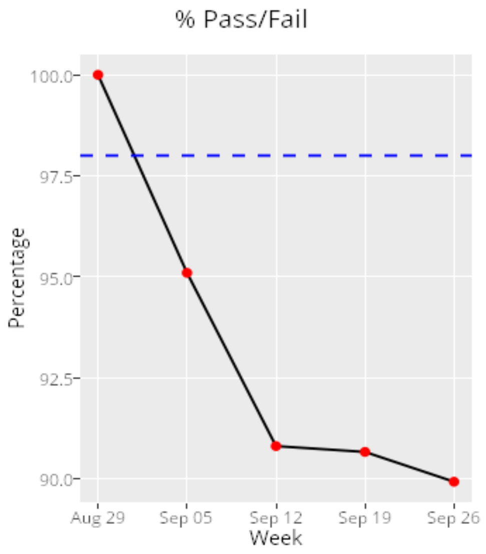


Figure 8: Distribution of Shipment Pass/Fail

Figure 9: Percentage of SLA Met per Week

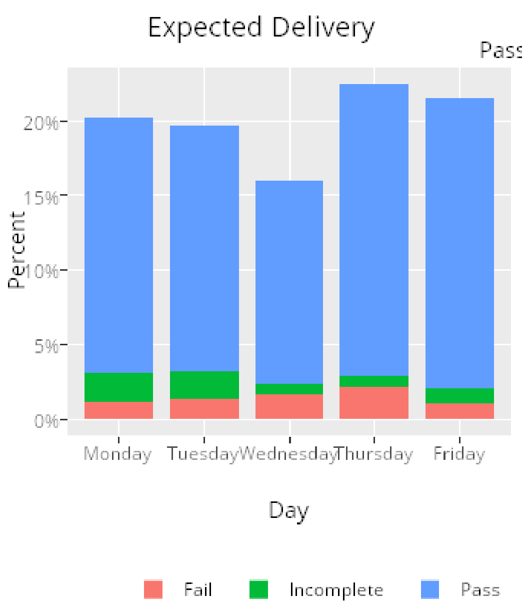
On the summary tab, 2 graphs are created to show an overview of the statuses of all the shipments for each country. These graphs have been filtered to either Inbound or Outbound shipments for clearer understanding of the shipment progress.

Figure 8 shows the distribution of all the shipments with regards to its status being “Pass”, “Fail” or “Incomplete”. The term “Pass” is referred to as having a turnaround time lesser than or equal to the stated SLA in our project. On the other hand, “Fail” would refer to having a turnaround time exceeding the SLA, and “Incomplete” refers to a shipment having a starting point but without any ending points hence cannot have its turnaround time calculated. This graph would allow our sponsors to easily understand the overall situation in a particular country, and to easily delve into shipments which have failed or are incomplete.

Figure 9 shows the percentage of shipments which have “Passed” for each working week. The x-axis shows the starting date of each week included in the data. The SLA threshold level is taken from an input, which allows our sponsors to tweak and adjust to the value desired instead of taking a static value. This graph thus gives a time series breakdown of the situation in each country.

1.2.2 DEEPER UNDERSTANDING OF THE SITUATION

The second tab of graphs provide a deeper understanding of the shipments in a specific country in providing more details on the shipment distribution performance.



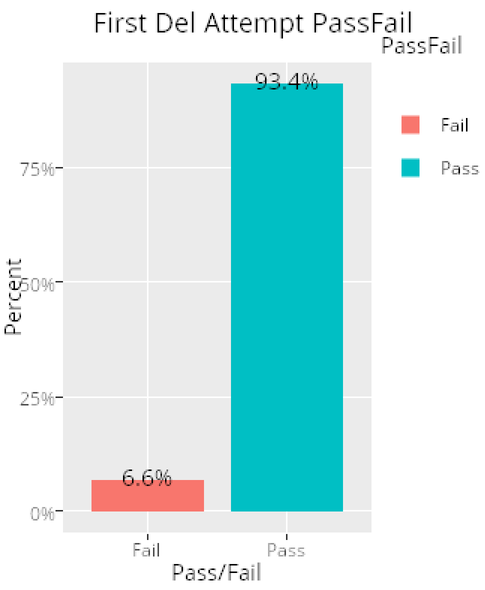


Figure 10: Expected Deliveries per Working Day

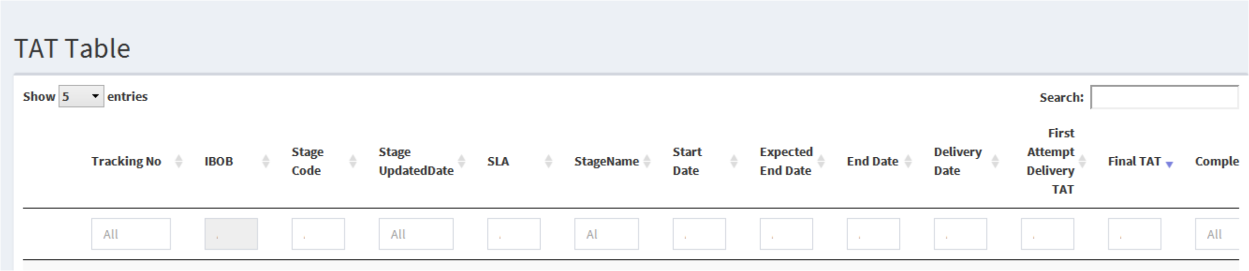
Figure 11: First Delivery Attempt Pass/Fail

Figure 10 shows the distribution of shipments for each working day denoted by the day of the week. This graph represents the expected first delivery day for each shipment. The expected arrival day is calculated by adding the starting date and the SLA number of days. If a shipment is labelled as “Pass”, this suggests that shipments have either arrived before the expected working day or arrived punctually on that day. On the other hand, if a shipment is labelled as “Fail”, it suggests that the shipment have arrived later than that day, and “Incomplete” suggesting that the shipment has yet to be completed. This graph allows our sponsors to expect and determine the shipments which should have arrived on a particular day but have failed, and thus delving deeper into the reasons for failure.

Figure 11 allows our sponsors to understand if shipments which have failed their first delivery attempt had completed the delivery eventually within the SLA. This allows our sponsors to easily identify shipments which have not completed their deliveries within the stipulated SLA.

To prepare the data for visualisation, numerous packages in R have been used. Firstly, *ggplot2* which is a popular plotting system used in Python and R for making professional looking plots have been used to create and display different graphs. Additionally, *plotly R* allows for making interactive quality graphs which have helped to create tooltips upon hover as well as create drilldown charts and tables for further insights.

1.2.3 TABLE



TAT Table

Show 5 entries

Search:

Tracking No	IBOB	Stage Code	Stage UpdatedDate	SLA	StageName	Start Date	Expected End Date	End Date	Delivery Date	First Attempt Delivery TAT	Final TAT	Comple
All	.	.	All	.	Al	.	.	.	.	.	.	All

Figure 12: Screenshot of Data Table in Dashboard

The data columns used in this data table were chosen for their relevance to the performance measurements used by the sponsors. The key data recorded in the systems were included in the data table, along with new data columns computed.

The calculation of the turnaround time (TAT) which allows for flexibility with the sponsors in the selection of endpoints and delivery statuses to be taken into account is the distinguishing factor of our data table compared to other methods currently employed by our sponsors. The data table recalculates the TAT based on the selections of the statuses by the sponsors, and also shows the corresponding SLA for each tracking number based on the data the sponsors have provided for us. We have also included two types of TAT calculations, one for the first delivery attempt which is used to assess the performance of vendors, another is the Final TAT which calculates the total TAT from the start to the end point. Some shipments may see a first attempt delivery TAT within the SLA, however saw a rather long total TAT. By identifying such shipments, the sponsors may look into the data for more details such as the reason codes to see if there is a potential issue that may affect future shipments, or is it an isolated case relating to the individual consumer.

The data table also shows whether a shipment has passed or failed its SLA requirement, making it easy for the sales manager to check the details with a simple filter instead of having to compute and compare the data themselves to find out whether it passed or failed. Data pertaining to the geolocation of the different shipments have also been added on the data table, for that some locations that may not be easily accessed may see longer TAT, which might be a point for the sponsors to explain to the client that they might not be able to account for.

The data table uses the *DT* package for R Shiny, which provides an R interface to using the JavaScript library DataTables, creating R data objects that can be displayed as tables on HTML pages with other features for higher degree of manipulating the data tables.

In preparing the data for the data table, we have used the packages *dplyr*, *plyr*, *timeDate*, *bizdays* to perform data cleaning and calculations. Dplyr, in particular, allowed for manipulating data frames with operations like SQL functions which made it a lot easier in cleaning up the data and performing data table functions.

1.2.4 MAP

In order to perform accurate and meaningful geospatial analysis, we utilised the following R packages. *Tmap* is the thematic maps package which provides geographical maps in which spatial data distributions are visualized. This package was used with its key ability to create multiple flexible choropleth maps. The *tmaptools* package provides a set of tools for reading and processing spatial data. This package was utilised together with *tmap* to map data over into the relevant polygons. The *leaflet* package provided the base layer map, which is a basic geographical and visually pleasing map of the world. *Leaflet* also allows better usability for users as zooming in and out is enabled with the scrolling of the mouse. To write and save the shapefiles, the *rgdal* package allowed this to be done efficiently with straightforward methods. The publicly available online shapefiles were rather large to read into the application and caused a lot of loading issues. So, to improve the loading time, we used the gSimplify method from the *rgeos* package to reduce the quality of the polygons but still retaining the shape and accuracy.

The map of Japan and Australia are used to reflect the Percentage of Passes for both Inbound and Outbound across all Tracking Numbers. Two layers of choropleth mapping, one each for inbound and outbound, are used to represent the percentages across the entire country. This allows the sponsors to tell which areas are more crucial by the colour (Figure 13 and upon interacting with the app, the pop up will display the total number of passes and failures for that selected area (Figure 14).

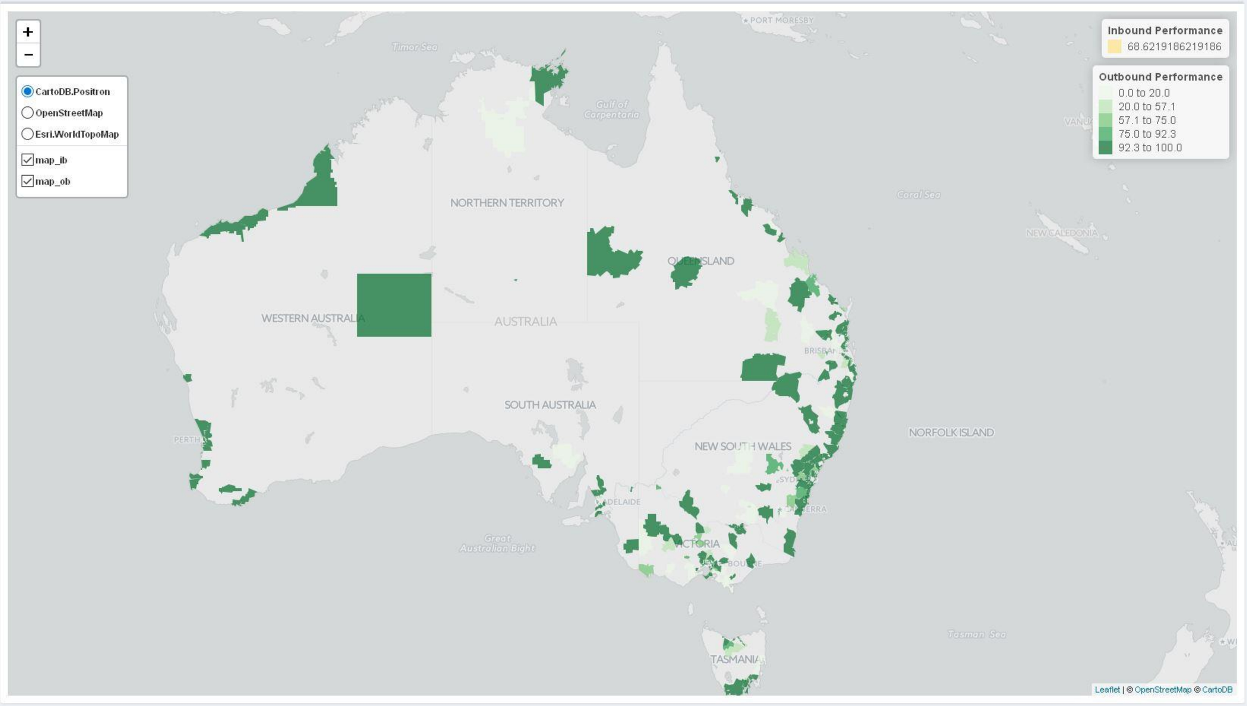
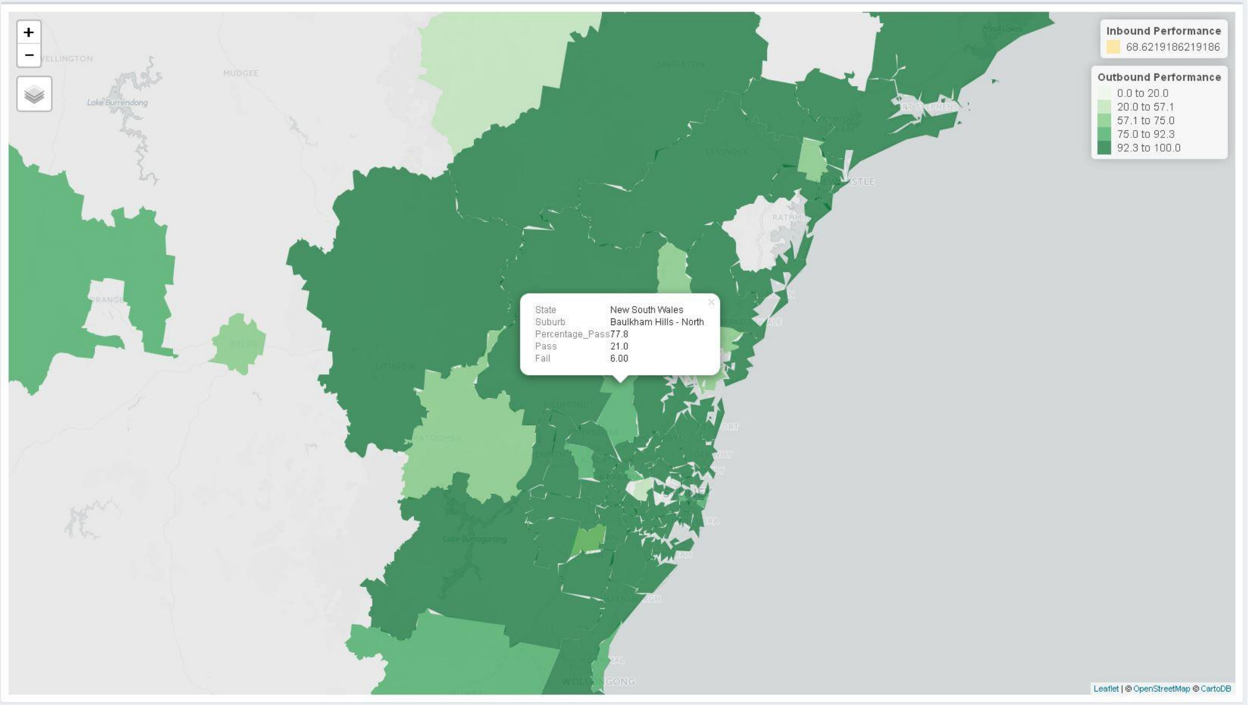


Figure 13: Australia's Map Level 1



State New South Wales
Suburb Baulkham Hills - North
Percentage_Pass 77.8
Pass 21.0
Fail 6.00

Figure 14: Australia's Map Tooltip

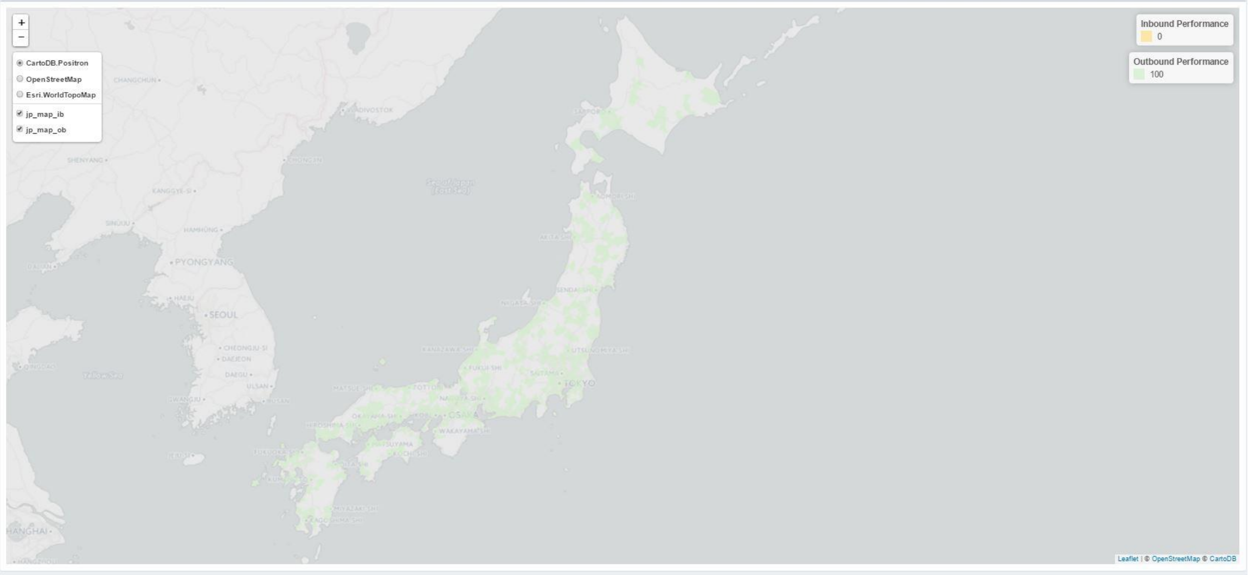
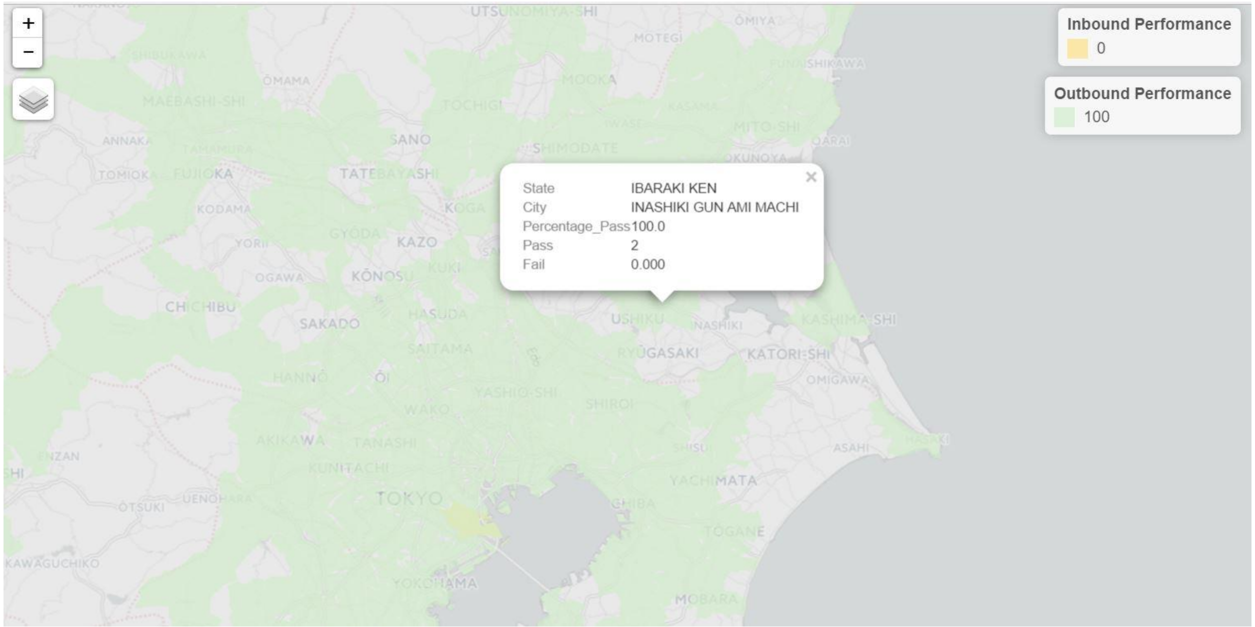


Figure 15: Japan's Map Level 1



State IBARAKI KEN
City INASHIKI GUN AMI MACHI
Percentage_Pass 100.0
Pass 2
Fail 0.000

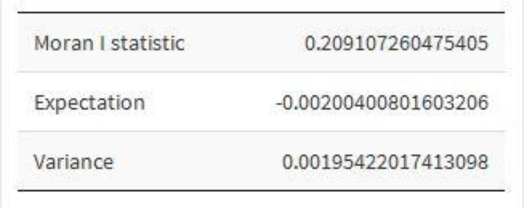
Inbound Performance
🟨 0

Outbound Performance
🟩 100

Figure 16: Japan's Map Tooltip

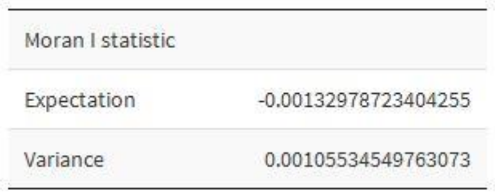
As shown above for the Japan data, the current dataset provided by the sponsors do not have any geospatial information pertaining to the Outbound shipments. Even so, the map segment of the application is able to cater in empty datasets or null values without issues and handle the sponsor's data with no problems.

On top of the two choropleth maps, we noticed that there are spatial patterns present and hence decided to investigate further. As seen from Figure 14, cities that have a high Percentage Pass tend to be surrounded by cities with high Percentage Passes as well. If this were true for Percentage Fails, the sponsor could potentially investigate the specific reasons for this. An example of a reason which might explain this pattern would be the courier responsible for delivering in those cities. To measure the spatial autocorrelation, we utilised Moran I. Moran I was carried out on the outbound shipments for both Japan and Australia.



Moran I statistic	0.209107260475405
Expectation	-0.00200400801603206
Variance	0.00195422017413098

Figure 17: Moran I Statistic for Australia Outbound Shipments



Moran I statistic	
Expectation	-0.00132978723404255
Variance	0.00105534549763073

Figure 18: Moran I Statistic for Japan Outbound Shipments

The results for Australia showed that there was indeed a spatial autocorrelation between cities (p-value > 0). However, the Moran I was not able to produce anything meaningful in the case of Japan because the data provided, which was only three months' worth, had a 100% Passing rate for all cities across Japan.

CONCLUSION

There is a need for greater sensitivity and delicacy in handling regional data, for that the data would also capture some nuances of the context it originates from, and is not to be assumed same across and handled without concern for it. With the globalization and increasing ties amongst world economies, we expect to see more of such data to be handled by future data analysts. Analysts working with the data need to broaden their horizons in understanding each context in aiding their understanding of the data to not mistake unfamiliar data captured as irregularities or errors.

Through this project, we learnt the complexities in the data across the regions and have learnt from our sponsors the ways to analyse data of this specific context in creating meaning and intuition for the sales team in charge. A data analyst needs to understand the needs and concerns of the sales team who will be using this data while fronting the customers, and also the concerns of the IT team who manage the systems to collect and transform this data collected. Our role in this project involves more than just working with the data given, but also to bridge the gap between the sales team and the IT team in providing suggestions from our understanding, both as data analysts and as consumers ourselves. As students trained in SMU learning both the general business curriculum and the analytics curriculum, it is important for us to demonstrate our multidisciplinary understanding of the data and apply it in our work. This is particularly the case as we have seen instances where there were communication gaps between the sales team and the IT team in conveying their requirements and needs.

Moving forward, we would wish to be able to build the application for the remaining 5 countries: Taiwan, Thailand, Hong Kong, New Zealand, and Korea. This would probably lead to purchases of national data from various countries. Currently 2 separate applications are built for Japan and Australia, moving forward we would seek to combine them both into a single dashboard, and then going ahead with incorporating all 7 countries. In doing so, we would have to ensure that the data load is not too much for the server to handle. While we did run into loading problems with the current free local host, we believe that when this application is adopted for commercial use and a better provider version.

REFERENCES

DuVander, Adam. (2010). *Map scripting 101: an example-driven guide to building interactive maps with bing, yahoo!, and google maps*. Chapter 3. [Books24x7 version] Available at http://common.books24x7.com.libproxy.smu.edu.sg/toc.aspx?bookid=41555

Goldberg, Swift, and Wilson. (2008). "Geocoding Best Practices: Reference Data, Input Data, and Matching Data." Page 5. Los Angeles, CA, University of Southern California GIS Research Laboratory Technical Report No 8

Goldberg, D. W., & Cockburn, M. G. (2012). "The Effect of Administrative Boundaries and Geocoding Error on Cancer Rates in California." *Spatial and Spatio-Temporal Epidemiology*, *3*(1), 39–54. Available at http://doi.org/10.1016/j.sste.2012.02.005

Ministry of Internal Affairs and Communications (Japan). "全国地方公共団体コード". Available at http://www.soumu.go.jp/denshijiti/code.html