# Introduction & Business Problem

With international traveling becoming more accessible throughout the world, deciding on where to go for your next trip is increasingly that of an arduous task. Where each convenience for which the intention is to give travellers a better understanding of potential destinations is followed by confusion if presented with too much information and too many options. Though detailed prospecting of destinations and potential planning cannot be avoided, alleviating the process with simple but relevant clues in areas such as relative density of landmark sites, child friendliness and cuisine could facilitate the final decision. Descriptive analyses could certainly be applied in order to help reach decisions on where to go and would most likely be the sole option for first time travellers as well as for those with very specific tastes. To alleviate the process, segmenting cities according to their venues characteristics around the city cores we can provide generalised profiles for prospective travellers to make use of when deciding on their travel destinations.

Once the travel profiles have been created, further application refinement could be developed by adding further data layers from either other FourSquare services internally or by connecting data from external API services such as TripAdvisor to our own as to add detailed information on for example identified venues, which in turn could lead to further refinements in our clustering of travel destinations.

## Main Audience

The main audience for the app will be **prospective travellers** **of all kinds** who will be given a choice of a set of segments of cities for their next travel destinations, characterised by venue and amenity availability rather than for example geography and popularity.

## Secondary Uses

Other use cases would be for **establishment owners** to use the differences between the clusters to identify business opportunities. Cities with a low hotel density could be interesting for **developers**. Up and coming expat cities will need amenities to cater for the influx of the new category of residents for which **city planners** can use data to pro-actively cater for relevant entrepreneurs to set up shop. In the same way, cities with ambitions to attract tourists may want to understand what world tourist cities make available for their guests.

## Data Sources

Find below the main data sources to be used in this exercise:

**Data Source 1 -** <https://www.worldatlas.com/citypops.htm>

The data list is a gross list of the largest cities in the world. Other criteria may be used such as most visited destinations to further find already, in a sense, empirically proven concepts. Doing so would however potentially filter away alternatives that may not have been “discovered” yet. The main parameters that we will be using are the city names for which we can use to extract the location co-ordinates for insertion into the Foursquare API calls. However, for the sake of completeness, we will also extract the population should we want for example to create per capita measures later on.

|  |  |  |
| --- | --- | --- |
| **Data Source** | **Feature** | **Description** |
| World Atlas Web Page | City | Name of cities to be used to extract the geolocations of the same. |
| World Atlas Web Page | Population | Population of cities in absolute count. Will be used as reference only and only included should there be measures that can be transformed using the values. |
| Geopy Function | Latitude | Location data to be matched with cities and to be used when extracting venues details later on. |
| Geopy Function | Longitude |

## Data Source 2 – Foursquare Venues API – Venue Details

Once the cities have been extracted from Data Source – 1 Foursquare Venues API to extract venues and their categories within the **maximum radius of 100,000 meters from the city core**. The features that will be extracted will be only 2 from the Venues API:

|  |  |  |
| --- | --- | --- |
| **Data Source** | **Feature** | **Description** |
| Foursquare Venues API | Name | Name of venue. It will not be used in the clustering but still good to extract to have as reference. |
| Foursquare Venues API | Venue Category | Level 1 venue category to be extracted. Level 2 category is a bit too detailed and will as such be left out of scope. |

**Source:** <https://developer.foursquare.com/docs/api/venues/details>

## Merged data set

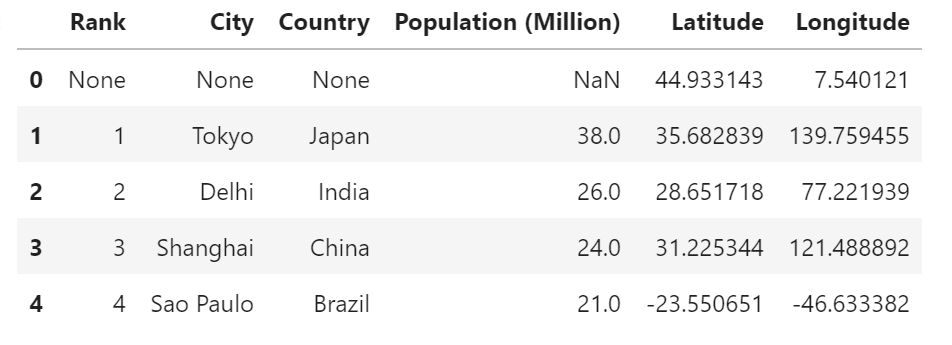
After extracting all relevant features we will be merging them into the below data frame for further analysis & clustering.

|  |  |  |
| --- | --- | --- |
| **Data Source** | **Feature** | **Description** |
| World Atlas Web Page | City | Name of cities to be used to extract the geolocations of the same. |
| World Atlas Web Page | Population | Population of cities in absolute count. Will be used as reference only and only included should there be measures that can be transformed using the values. |
| Foursquare Venues API | Venue Name | Name of venue. It will not be used in the clustering but still good to extract to have as reference. |
| Foursquare Venues API | Venue Category | Level 1 venue category to be extracted. Level 2 category is a bit too detailed and will as such be left out of scope. |
| Geopy Function | Latitude | Location data to be matched with cities and to be used when extracting venues details later on. |
| Geopy Function | Longitude |

## Data Sample Construction

### Cities Data Set

The data is extracted from <https://www.worldatlas.com/citypops.htm> and then matched with geolocations that we pull using the geopy function and put into the below data table. For now we also take note of the None city values for which we will have to clean later on.

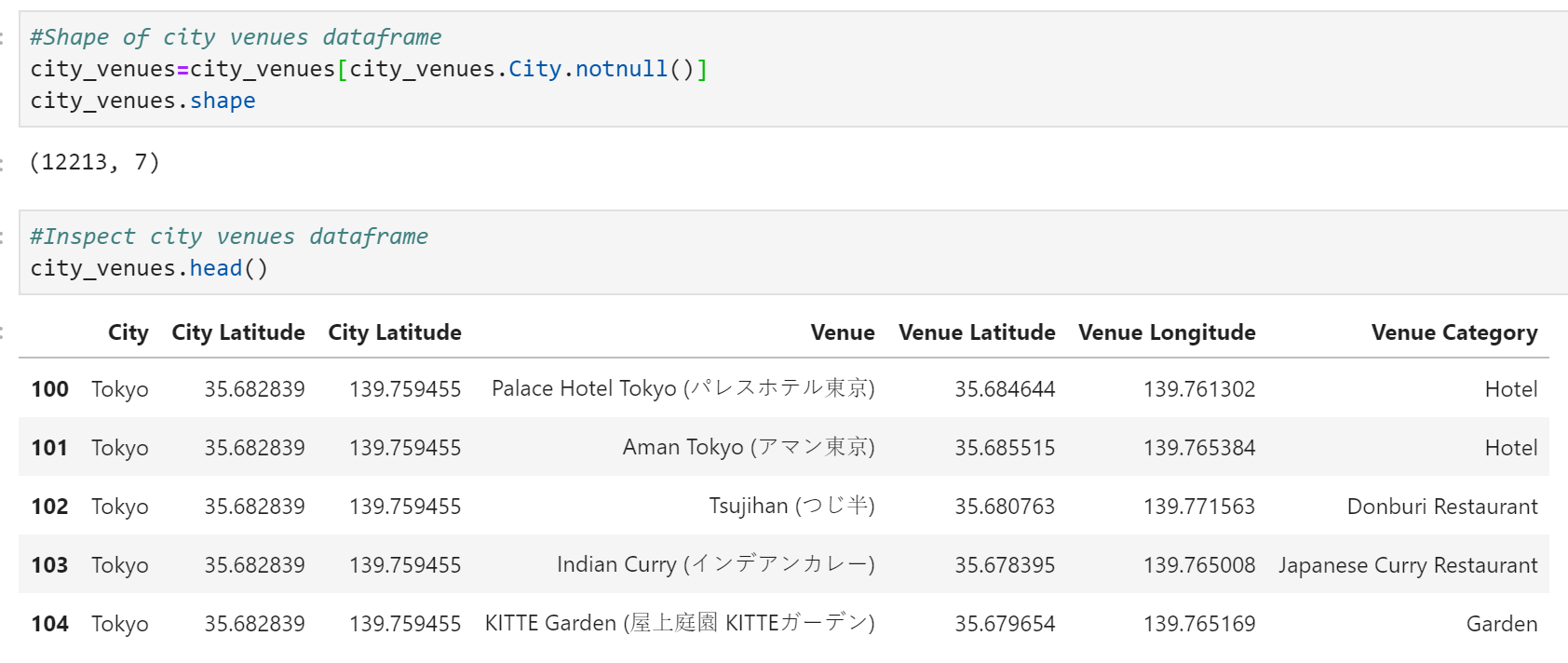


|  |  |
| --- | --- |
| **World** **Cities Plot** | |
|  | The figure represents the cities that we have scraped from the worldatlas.com website. Cities represented on the graph are the most populous in the world and are the chosen sample for which we will be extracting Foursquare venues from.  Overall, it seems like we have a good coverage across the globe and major cities from |

### Foursquare Data

#### City Venues Data Frame Construction

Using the city geolocations we call the Foursquare Venues API to extract their details, including our main variable of interest – ‘’Venue Category”. The resulting data frame once we have consolidated the can be seen as per the below snippet. We note that, after dropping any null values that may have followed into the construction of the data set, we have 12,213 rows of data in total to work with which should provide plenty of information to be able to pass onto our machine learning algorithms later on.



|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

# Methodology

## Exploratory Data Analysis

#### Venues Count

|  |  |  |  |
| --- | --- | --- | --- |
|  | City | Venue | Venue Category |
| Count | 12213 | 12213 | 12213 |
| Unique | 150 | 10965 | 511 |
| Frequency | 100 | 186 | 1046 |

Overall statistics of our final data set can be seen above. We can see that we have 150 unique cities as expected where each city [[1]](#footnote-1) has 100 venues tied to them (this is a technical limitation of the Foursquare API). We also see that we have a large number of venues spread across a total of 511 categories.

|  |  |
| --- | --- |
| **Data Set Venue Category Distribution** | |
|  | Taking a quick look at the data seem to confirm that our sample first and foremost is suitable to use in a travel destination recommendation engine as the by far most common venue type is hotels. Moving downwards in the list shows an overrepresentation of eateries and related venues. We do however also find. |

## Inferential Data Analysis

No classical inferential data analysis such as correlation, medians, averages, quartiles and so on can be readily used for our data set due to it comprising mostly of objects rather than numerical values. Therefore, the main bulk of the analysis will be done on our machine learning results for which we will need to interpret in order to make our qualitative recommendations.

## Segmentation – Most Common Venues



The basis of creating our city segments will be the relative frequency of venue categories of the cities. By counting and ranking them as per the above table, we have a data set ready to be used in our machine learning algorithms. For the project, experimentation was also made on the number of most common venues i.e. the number of columns were made. 10 was decided to be a good number of most common venues as it will add diversity to our results. Experimentations were made with a smaller number of rankings with the result being that a large amount of cities

## Machine Learning Method

An unsupervised **k-Means Clustering Algorithm** has been chosen to extract insights from our data set. The algorithm is used as we have no readily provided city characteristic labels that we can use. The end product is also self-contained in the sense that we will not attempt to classify further cities by entering new records; we only want to organise and understand our finite set of data in a cluster format as to provide end-users with travel destination segments based off existing, unstructured, characteristics. Though arbitrary, **we choose 5 as the number of Ks** when running our machine learning algorithm, the rationale being that too many will confuse travellers and risk a dilution of results while too few won’t give us a satisfactory diversification.

# Results & Recommendation

Below is a description of the 5 segments generated by the clustering algorithm. These can be viewed directly in our python notebook & Excel file available for download in our GitHub directory. To further understand our clusters, we have opted to analyse them from the perspective of how many times each category types appear in the top 10 most common venues of each city in percentage. Doing so visualises the decision criteria that the clustering algorithm used for us to make sense of and interpret the results. The metric was chosen above the absolute count as our resulting segments include different number of cities for which pure counts wouldn’t be comparable. A detailed list of cities and what cluster they belong to can be found in the Appendix.

|  |  |
| --- | --- |
|  | The coloured dots on the world map represents the 5 clusters that we have assigned the 150 scraped cities to, based off the 10 most common venues present in each one of them.  The segments seem to have an even spread throughout the world which is a promising result as this will give travellers a broader selection of similar cities to choose from depending on their individual preferences regarding what kind of a holiday that they are looking for. |

## City Segments

Find below the characteristics of the segments together with a brief qualitative description on their strengths for travellers to understand when picking his/her destination.

### Cluster 1 – City Shopping & Hangouts

|  |  |  |
| --- | --- | --- |
|  | | Our first segment has been named City Shopping & Hangouts due to the high frequency of shopping malls in combination with cafes and coffee shops in the top 10 most common venues lists. This segment of cities suits those looking for a convenient experience with lots of well needed opportunities for breaks between the last shop to the next while carrying around on bags.  **The segment consists of a total of 10 cities, listed in the Appendix section.** |
| **Venue Category** | **Appearance in 10 most common venues** |
| Hotel | 90% |
| Café | 60% |
| Shopping Mall | 60% |
| Bakery | 50% |
| Coffee Shop | 50% |
| Ice Cream Shop | 30% |
| Plaza | 30% |
| Italian Restaurant | 30% |
| Indian Restaurant | 30% |
| Fast Food Restaurant | 30% |

### Cluster 2 - City Hangout, Dining & Nightlife

|  |  |  |
| --- | --- | --- |
|  | | This segment is mainly characterised by the high frequency of cities with high densities of eateries of all kinds. Coupled together with the relatively high occurrence of bars in the top 10 venue types, these cities provide a well-rounded experience focusing more on the culinary and social aspects.  **The segment consists of a total of 38 cities.** |
| **Venue Category** | **Appearance in 10 most common venues** |
| Hotel | 74% |
| Café | 66% |
| Coffee Shop | 53% |
| Park | 39% |
| Ice Cream Shop | 37% |
| Shopping Mall | 37% |
| Italian Restaurant | 34% |
| Pizza Place | 32% |
| Restaurant | 29% |
| Bakery | 24% |

### Cluster 3 – Green City Hangout & Dining

|  |  |  |
| --- | --- | --- |
|  | | With a high rate of parks and eateries, these cities provide visitors with a city stay complemented with plenty dining opportunities in scenic environments thanks to the high frequency of parks.  **The segment consists of a total of 41 cities, listed in the Appendix section.** |
| **Venue Category** | **Appearance in 10 most common venues** |
| Hotel | 80% |
| Coffee Shop | 63% |
| Park | 46% |
| Café | 44% |
| Italian Restaurant | 39% |
| Bakery | 34% |
| Shopping Mall | 34% |
| Pizza Place | 32% |
| Restaurant | 32% |
| Ice Cream Shop | 24% |

### Cluster 4 – Shopping, Sweets & Nightlife

|  |  |  |
| --- | --- | --- |
|  | | The Shopping, Sweets & Nightlife segment provides an experience focused on indulging yourself that new jacket while enjoying local pastries to satisfy your sweet tooth. Once ready, there is also ample of opportunities to grab a fast food burger or fried chicken before heading out into the night.  **The segment consists of a total of 28 cities, listed in the Appendix section.** |
| **Venue Category** | **Appearance in 10 most common venues** |
| Hotel | 82% |
| Coffee Shop | 71% |
| Shopping Mall | 46% |
| Park | 43% |
| Café | 43% |
| Fast Food Restaurant | 39% |
| Bakery | 39% |
| Ice Cream Shop | 29% |
| Pizza Place | 21% |
| Bar | 21% |

### Cluster 5 – Green City, Shopping & Nightlife

|  |  |  |
| --- | --- | --- |
|  | | Want to enjoy the shopping & bubbling nightlife of cities without losing touch with nature? Look no more! This segment appeals to those who wants to strike a balance between city life and greenery.  **The segment consists of a total of 33 cities, listed in the Appendix section.** |
| **Venue Category** | **Appearance in 10 most common venues** |
| Hotel | 79% |
| Coffee Shop | 70% |
| Park | 61% |
| Shopping Mall | 48% |
| Café | 39% |
| Bar | 36% |
| Fast Food Restaurant | 30% |
| Restaurant | 27% |
| Pizza Place | 27% |
| Italian Restaurant | 24% |

## Summary & Results Notes

By clustering the cities according to exhibited characteristics, we receive 5 segments that we can present to travellers that are still deciding on where to go. Want good dining opportunities as well as ample of opportunities to continue out afterwards? Choose cluster 2. Want the same with without the nightlife but instead a somewhat calmer atmosphere in order to have that romantic walk through the park back to the hotel? Choose cluster 3. Once a segment has been chosen the traveller can then decide on where to go given the

Despite the clustering algorithms results returning 5 segments, looking at the detailed results tells us that while there are differences, there may be more similarities than dissimilarities between the 5. This is not surprising as large cities tend to develop similarly in terms of amenities & venue types, with hotels, parks and eateries being some of the top categories that cater to both visitors’ and inhabitants’ base needs. We therefore recognise the need to have a further discussion on what can be done in terms of improving our results, whether it be augmenting our data set in order to be able to control our result, overcome data sample limitations, sample & feature biases or further massaging and transforming our data.

# Discussion

## Recommendations

### Travelers

With our constructed segments travellers can choose to use the insights from 2 different perspectives:

1. **Travel Suggestions**Travellers can choose from a list of desired city trip experienced as defined in our previous Results section. Once chosen, the user will then then get a list of cities to choose from that fall into the preferred category before proceeding to booking the ticket or choosing to do more research. In the same way, users can also enter a city name and then be returned with what kind of experience that can be expected from going there. This is handy when the user receives a recommendation without too many specifics on what to expect.
2. **Finding similar destinations**In the cases where a user wants to have an experience similar to a past trip but to a new destination, he/she can enter the city name and be returned with the full list of cities contained within the cluster of belonging.

### City Planners, developers & entrepreneurs

In the same manner but opposite direction for the above, city planners, developers and entrepreneurs can make use of the insights to identify business opportunities to cover the weak points of the cities. Is a city strong in shopping and dining but lack green areas? Build more parks! Does the city die after 7PM after the last shop closes? Set up bars!

## Methodology refinements

In this project we have applied an unsupervised learning algorithm. While not possible with the current data set, one can start to manually label the cities given their top venues. Once a satisfactory testing data set has been assembled, a supervised learning algorithm can leverage on the new information and classify the rest of the sample or accept external input of new cities and characteristics. This method is preferable if we desire to control exactly how many segments we want to be used and what they should be defined as, versus the current unsupervised procedure where we uncover hidden relationships within a set of unlabelled data.

## Sample data limitations

|  |  |
| --- | --- |
|  | The data used in this study has some limitations given our scope that will have to be taken into regard during the inference of our results. For one, the venues limit per city is 100 which coincides with the API call limit. Counting the values per city will confirm the fact as per the screenshot provided. With the limit in place, we will not get a true representation of the cities’ venues.  In the same way, for cities with a very low number of venues, venues may be missing from the Foursquare API which limits any inferences that we can make using the date. |

### Sample bias - Cities

The city list used as our sample is based off population. While population may be an indicator of suitability of the travel destination, it will also miss out on smaller cities population wise that may be of interest such as Stockholm with its roughly 1.5 million inhabitants. The city itself may be a bit smaller but is worthy to be of consideration for any traveller with its rich history, beautiful nature and wide selection of amenities & entertainment options. In the same way, the data set can also be skewed in favour of population rich countries such as China and India. While not always true, high population countries may also contain a larger number of large cities where not all of them may be of interest for international travellers to go to.

## Omitted feature bias & filtering

Our chosen main variable of interest, venue category, may not be enough in order to fully segment our cities to suit our purpose of providing accurate travel recommendations. One example would be adding **region of the world** and run our algorithms for each one of them. This is relevant in the case where travellers have already chosen a general area of the world to go to such as South East Asia, and only want recommendations from that part of the world.

Another measure would also be adding the **destination popularity** using for example international visitor statistics. This will add another potential filter where prospective travellers can choose well established tourist destinations or go for less popular sites.

**Price Levels** in general can also be added to further help the user on where to go. This is, for example, especially important for travellers from developing countries as the price levels are in many cases much higher in the developed parts of the world.

## Feature data refinement

Currently, all venue categories that are available in the Foursquare API are used in our machine learning algorithms without any transformations. However, given our results, performing further operations on our dataset may be desirable to even out some of the kinks of our results.

* **Eateries**  
  The Foursquare venues details data set provides a detailed categorisation of eateries. While useful, in this application it would perhaps make a bit more sense to consolidate outlets that sell food to a smaller number of categories such as local or international restaurants. Doing so will still assign weight to how much
* **Cultural venues – shifted weighting & groupings**

While eateries are overrepresented, the results indicate that the cultural venues such as museums and art galleries are not given enough weight in the clustering algorithm. Finding a way to either consolidate these venues or modify their weighting will allow for a new set of segments characterised by their cultural merits.

# Conclusion

Scraping the worldatlas.com top 150 most populous cities and clustering the same gives us 5 clusters that gives undecided travellers material to use when deciding on where to go next. The five clusters inherently represent unique experiences that travellers can take a stance towards. Once the desired experience / outcome has been chosen, a list of cities will be presented to the user to pick from as a potential destination for his/her/their next city destination. City planners, developers and entrepreneurs can in turn use the same information in planning on the future path of cities and business opportunities to cover for any weak points that may be the case.

Though we can see what makes each segment unique, we also recognise that there may be more similarities than dissimilarities between them which will need to be addressed in future iterations of the application as part of the virtuous data science project cycle where insights from feedback loops is used to further improve on the product.

# Appendix

## List of cities & their segments

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Cluster 1** | **Cluster 2** | **Cluster 3** | **Cluster 4** | **Cluster 5** |
| Delhi | Mumbai | None | New York | Sao Paulo |
| Cairo | Mexico City | Tokyo | Karachi | Osaka |
| London | Beijing | Shanghai | Shenzhen | Istanbul |
| Hyderabad | Dhaka | Lagos | Jakarta | Guangzhou |
| Xi'an | Buenos Aires | Rio de Janeiro | Chennai | Los Angeles |
| Dar es Salaam | Kolkata | Paris | Nagoya | Moscow |
| Qingdao | Chongqing | Lima | Tehran | Kinshasa |
| Ji'nan | Manila | Bogotá | Nanjing | Tianjin |
| Detroit | Bangalore | Bangkok | Ho Chi Minh City | Seoul |
| Lucknow | Johannesburg | Wuhan | Baghdad | Lahore |
|  | Chicago | Chengdu | Santiago | Dongguan |
|  | Madrid | Ahmadabad | Riyadh | Foshan |
|  | Miami | Hong Kong | Belo Horizonte | Houston |
|  | Pune | Kuala Lumpur | Dallas | Suzhou |
|  | Barcelona | Hangzhou | Singapore | Haerbin |
|  | Ankara | Shenyang | Alexandria | Guadalajara |
|  | Zhengzhou | Toronto | Kunming | Yangon |
|  | Boston | Surat | Hefei | Kabul |
|  | Melbourne | Philadelphia | Fuzhou | Chittagong |
|  | Brasília | Kitakyushu | Shijiazhuang | Monterrey |
|  | Phoenix | Luanda | Nanning | Sydney |
|  | Fortaleza | Atlanta | Busan | Dalian |
|  | Changsha | Khartoum | Milan | Xiamen |
|  | Recife | Saint Petersburg | Mashhad | Jiddah |
|  | Rome | Washington, D.C. | Puebla | Shantou |
|  | Zhongshan | Abidjan | Sana'a | Tel Aviv |
|  | Cape Town | Montréal | Santo Domingo | Kano |
|  | Porto Alegre | Nairobi | Guatemala City | Berlin |
|  | Salvador | Medellín |  | Taiyuan |
|  | Faisalabad | Changchun |  | Addis Ababa |
|  | Aleppo | Hanoi |  | Wenzhou |
|  | Dakar | Casablanca |  | San Diego |
|  | Urumqi | San Francisco |  | Athens |
|  | Curitiba | Seattle |  |  |
|  | Jaipur | Ibadan |  |  |
|  | Shizuoka | Yaounde |  |  |
|  | Ningbo | Wuxi |  |  |
|  | Campinas | Izmir |  |  |
|  |  | Kanpur |  |  |
|  |  | Douala |  |  |
|  |  | Kiev |  |  |

1. For further comments and discussion, see section 4.3 [↑](#footnote-ref-1)