On the move

An exploration of population movement in a dynamic world and its cities

Word count: 6,337

Website link | GitHub repository link

List of Figures and Tables

1.0 | Introduction

Summary of our work and what this document achieves

This report outlines our approach to exploring and visualising population movement in a dynamic world and its cities. It speaks to the theme of *Living Cities* whereby we equate living with the dynamism of movement, specifically movement of people in the context of migration. This research is separated into two parts, international and national with China as a case study. We aim to satisfy both research and technical documentation criteria in order to allow for a greater understanding of design and analysis choices for our website which accompanies this report.

Movement of people, inseparable from systemic changes of globalisation, is one of the key traits of modern states and cities with an estimated 175 million people migrating to another country for living in the year 2000. Today, at the country level, the tension between international migration and growth of national interests often calls for the strengthening of international cooperation (Doyle, 2004). However, cities are not isolated, they are holding the majority of populations in countries and are the principal nodes of national and transnational migration networks. As a result, we focus on this interconnectedness that defines the places people inhabit.

Based on foundations in the literature, we use Allen's (2018) representation of the eight-dimension framework of migration, with amendments made in response to the data available. These come down to the emphasis of quantitative, temporal and spatial dimensions, however we do recognise that for a truly holistic approach all ought to be taken into account.

Two pairs of visualisations are generated utilising two distinct datasets, adapting existing work as well as create new bespoke visualisations and overcoming a number of compatibility, performance and visualisation challenges.

The output described in this document serves as an initial output in itself but equally a reproducible process which could be rerun in order to broaden the scale of analysis into additional time periods, thus providing a more holistic view of international and national population movements.

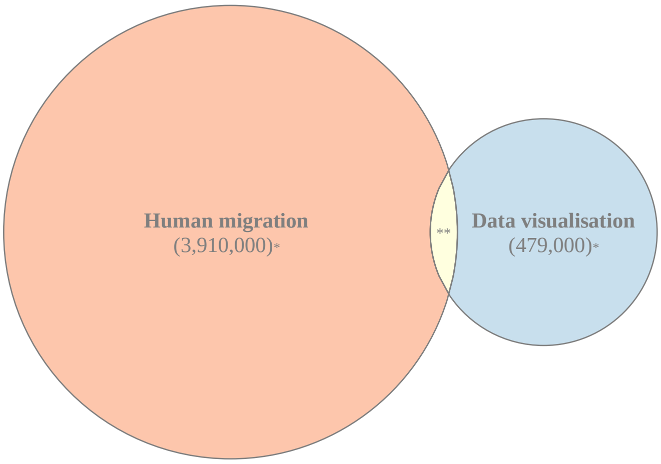
Overall, international migration continues to be heterogeneous among countries while maintaining a steady system-level intensity, aligned with expectations. However, beyond the aggregate stability, a shift can be observed, supported by theory in the literature with respected to the shifts of key migrants' origins. Furthermore, investigation of national-level dynamics in China using mobile phone generated data further supports the work in the literature carried out on lower temporal resolution datasets which are updated on an annual or lower frequency basis, thus suggesting that it has scope to act as a potential proxy if carefully prepared.

2.0 | Literature review

Summary of key literature on (inter)national migration and visualisation approaches

For this project, we considered two strands of literature, expansive in their own right, to appropriately situate our work and refine our research focus in a complimentary manner. Examined separately, studies regarding migration and visualisation are prohibitively broad. However, an overlap exists that combines the two strands where we aim to contribute with our work, as approximated by Figure 1. The size of segment '**', representing just over 50,000 results, suggests that while niche effective visualisation is an underutilised tool.

Figure 1 | Google scholar search results for project themes



**Note*: values indicate google scholar search results on 21st May 2020, size of shapes is representative.*
***Note*: overlapping segment value is 52,200.*

The below sections outline the key building blocks from the two strands of respective literature and their intersect followed by the statement of our research questions and hypotheses.

2.1 | Migration research literature

To structure our understanding and exploration of human migration, we turn to the eight-dimension framework outlined by Allen (2018) and recreated in Table 1.

Table 1 | Eight dimensions of migration and features (Allen, 2018)

Dimension	Description	Example features
1. Quantitative	Numerical data showing quantities of migrants	a. Migrant stocks **b. Migrant flows**
2. Type	Different reasons or main motivations for migrating	a. Labour/economic b. Forced c. Study d. Family
3. Spatial	Geographic origins, destinations or transiting areas	**a. Direction** **b. Locations** **c. Internal movement** d. **International movement**
4. Temporal	Changes in characteristics relating to individuals or migration processes over time	**a. Dynamics/trends** b. Histories c. Shocks
5. Political	Governance of migration or migrants via expression of power	a. Public opinion b. State activities c. Differences in policies
6. Social	Interpersonal and cultural aspects	a. Lived experiences b. Community impacts
7. Economic	Impacts and drivers of migration relating to labour markets or fiscal performance	a. Access to employment b. Impacts on public services
8. Ethical	Trade-offs, norms, values or moral dilemmas	a. Availability of rights b. Available freedoms

***Note**: original table can be found on page 6 of Allen (2018) under the title "Table 1. Dimensions of migration and corresponding features"*

It is important to recognise the difference between migrant stock and flows (Parsons et al., 2007).

*"**Migration flow** data may be compared with a **video recording** of every migration event during a given period, while **population stock** data provides **photographs** of a given population at a given moment." (Poulain, 2008, p. 45)*

This distinction in terms, takes on an added importance in certain national settings such as China, which as a result of its *hukou* system, affecting residency rules, has a definition and *hukou*-compliant migrant flow but also a so-called 'floating population' which is better as approximated as internal migrant stock which although on the move/'floating' is often 'statistically invisible' (Chan, 2013). Therefore, an appropriate choice of dataset is required when examining these dynamics. In keeping with the movement-centric theme of this project, we will focus on feature '1b' and the flow of migrants and observable trends in their flow. This will involve *any* population on the move across geographical levels of analysis.

Dimension 2 speaks to the importance of categorisation of migration linked to its principal motivation. However, although necessary to acknowledge and to truly understand the dynamics behind the flow of migrants, the availability of data is historically inconsistent (Parsons et al., 2007). Most recently, the work of Tjaden et al. (2019) reinforces the unchanged nature of this fact as is it is itself limited to an intention to migrate, rather than a clear driver of that intention. As a result, it is a lower priority consideration when considering global or aggregated national flow datasets and will be out of scope for this project.

In understanding flows, dimensions 3 and 4 are critical in that they capture spatial-temporal dynamics. The emphasis is on capturing the origin and destination of a migrant across a national (internal) or international scale across time to understand broader systemic trends. Some observed trends include the increase of lifetime migrants increasing by ~ 2.2% annually between 2000 and 2015 (Edmonston and Lee, 2018). However, in line with the work of Vezzoli (2017), care must be taken when taking historical trends and projecting them into the future.

A general consensus can be observed in that economic opportunity is a fundamental driver of migration, be that nationally or internationally (Chan, 2013; Vezzoli et al., 2017; Zhang, 2017). With the correct datasets, work can be done across trends in origin and destination but also on trends over time and seasonality when data is available at higher temporal resolutions.

The remaining four dimensions, as dimension 2, are important but are beyond the scope of this paper to gather on a consistent basis and at scale for high-level datasets. Therefore, although acknowledged as important dimensions to be discussed, they will not be explicitly taken into account for the purpose of this exploratory visualisation piece of work.

2.2 | Migration visualisation literature

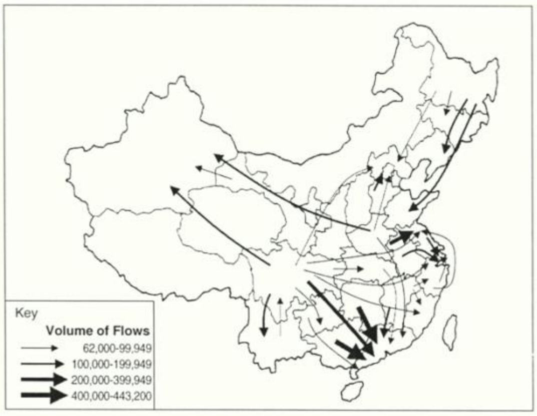
An understandable, but pertinent, theme throughout the literature reviewed above is the absence of visual storytelling to reinforce the arguments made or make the data at hand more accessible. Below we outline key works and approaches that inform our work and shaped our research agenda.

At an international level, Sander et al. (2014) stands as a differentiated piece of work bringing international migration and visualisation disciplines together to explore a central phenomenon of a globalised world in a clear and succinct manner. At the heart of their work is bespoke estimation process for bilateral migration flows utilising United Nations data. Further, results are presented in 5-year windows from 1990 to 2010 using a ‘circular plot’ (Abel and Sander, 2014). Subsequent to publication, Abel and Sander’s work has been criticised for deviation from the framework in Table 1 (Allen, 2018; Fakir and Abedin, 2020; Kennedy et al., 2016; Vannini et al., 2018). Nevertheless, it stands as an important illustration of the junction of the two disciplines that has notably not been updated since its publication in 2014 prompting us to fill this gap.

At a national level, migration and visualisation appears to be slightly ahead of the international level, however the dataset used are often confined to datasets at a lower temporal resolution (annual or longer). Given the extensive movement of peoples within its borders, we use China as a bellwether by examining select works from the past 20 years. We can observe the use of smaller datasets and showcase the appropriate but narrower range of visualisations used below with some works omitting spatial visualisation altogether (Goodkind and West, 2002; Hare, 1999). This has encouraged us to experiment with alternative datasets to test whether patters such as those shown in Figure 2 and Figure 3 continue to be observed.

Figure 2 | Visualising internal migration flow aggregated at province level (Chan, 2001)

Figure 1. The Largest 30 Inter-provincial Migration Streams, 1990-1995



Source: National Population Sample Survey Office, *1995 Quanguo 1% Renkou Jiuyang Diaocah ziliao (Data on 1995 National 1 percent Population Sample Survey)* (Beijing: Tongji Chubanshe, 1997).

Note: *Figure 2 showcases the core movements of China's labour market, using official statistics and discerning key flow directions and magnitudes. Of note is additional work, such as that of Wu and Zhou (1996) that argues for the economic rationality of such moves and the fact that trends observed are no different from other economies. The implication is that with the standardisation of data captured, same techniques can be used to visualise flows such as these across other countries where a gravitational-like pull would be expected towards the areas of high economic activity.*

Figure 3 | Visualising province level observations

Figure 3A | Net migration and GDP (Chan and Wang, 2008)

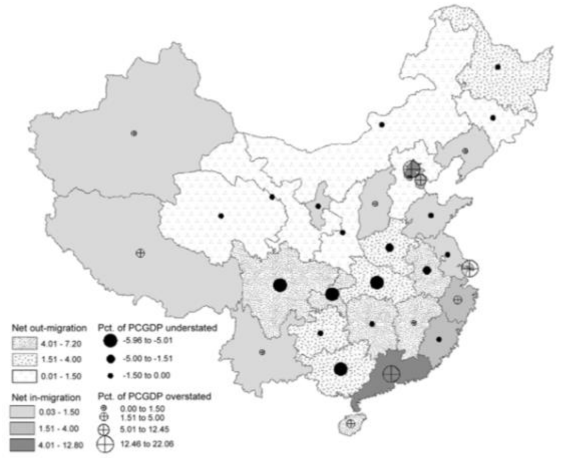
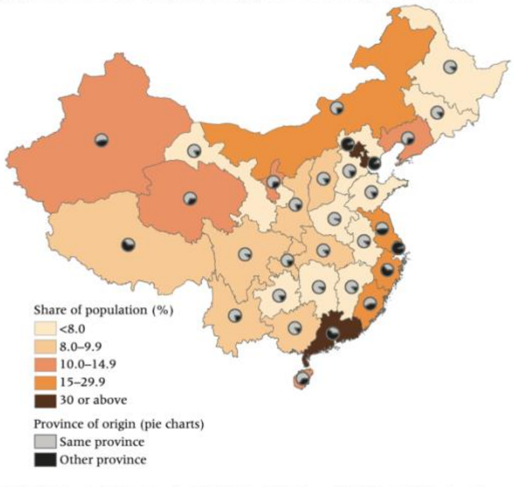


Fig. 3. Net migration (1995–2000) and over/understatement of per capita GDP in 2000 by province. Net migration is net interprovincial migration between 1995 and 2000 expressed as a percentage of the *de facto* population in 2000. Over/understatement of per capita GDP is the difference between the GDPs based on the *de jure* and *de facto* populations, expressed as a percentage of the *de facto* population in 2000. Compiled by the authors from NBS (1996, 2001, 2007a) and SC and NBS (2002).

Figure 3B | Floating populations (Liang et al., 2014)

MAP 1 Inter-county floating population as a share of provincial population and its composition by province of origin, China, 2010



SOURCES: Data for 2010 are based on PCO (2012), Table 7-1, pp. 458–459 and 1/1000 micro-data.

Note: *As per the above, Figure 3 summarises data at province level and uses datasets at annual resolution. Notably, Liang et al. (2014) indicates two important changes in movement patterns. Firstly, the rising importance of intra-province migration relative to inter-province migration. Secondly, a shift in destination for inter-provincial migration toward the Yangtze River Delta (surrounding Shanghai) instead of southern China (around Hong Kong, Shenzhen, Guangzhou).*

2.3 | Research questions

To summarise, two overarching questions are noted above.

Research question 1 | Using the visualisation techniques of Sander et al. (2014), do we see comparable migration trends continue in the decade since the original publication?

This has an implication on our understanding of migration as well as a further standardisation of open data that could be refreshed regularly and provide a baseline of magnitude through which to contextualise a country's own migration, as well as contrast the numbers of national (internal) and international migration respectively.

Research question 2 | Given the proliferation and access to alternative data sources, can mobile positioning data be used as a proxy dataset for internal migration?

> **Research question 2.1 |** Do the daily movements of people follow the *direction* and *intensity* suggested by annual statistics from official sources?
>
> **Research question 2.2 |** Can any seasonal patterns we discerned from the data?

The implication of the work is potentially interesting as it could open more detailed avenues for research as well as better indicator of which cities are likely to be 'winning' or 'losing' from migration (Ma and Tang, 2020).

3.0 | Data and methodology

Summary of our data sources and our data processing

This section outlines key aspects of data sources to address our research questions and generate visualisation outputs. As per Section 2, our focus across two levels demands two disparate datasets.

Firstly, we diverge from Sander et al. (2014) and their bespoke estimation technique to utilise data from the United Nations Population Division in its existent form and is broken down in Section 3.1. Secondly, we have procured data from the Tencent Data Store with aggregated mobile phone information with a degree of 'heat' which in effect represent the intensity of inter-city population movements and can be loosely used to approximate the number of total travels (which may not correspond to individuals). This is elaborated on in Section 3.2.

3.1 | International migration data – UN DESA (Population Division)

This dataset from the United Nations Department for Economic and Social Affairs and its Population Division. It stores origin-destination data for migrants at an annual level on a country to country basis. In effect, the excel, accessible for years 1990 – 2019, is a matrix (270 columns by 270 rows) per year of countries and regions and the movement between them measured by individuals. Further breakdowns are available such as those by gender, however those, although relevant to a number of migration dimensions (as per Section 2), will be omitted in order to maintain a focused scope.

Table 2 | Illustration of UN migration dataset

Population outflow	Western Asia	Central Asia	Southern Asia	Eastern Asia	Eastern Europe	Northern Europe	Southern Europe	Western Europe
Western Asia	13,808,956	95,089	18,402,429	32,939	1,326,557	179,540	331,077	596,121
Central Asia	166,412	482,760	16,363	102,415	4,395,501	11,502	0	11,699
Southern Asia	161,692	11,811	11,176,671	312,886	1,473	54,939	5,670	5,030
Eastern Asia	235	28,041	210,013	5,201,972	21,384	60,989	11,016	35,205
Eastern Europe	2,097,067	5,603,254	1,666,384	830,016	10,339,711	489,026	389,687	426,144
Northern Europe	1,135,485	41,672	2,550,790	580,099	2,472,166	2,050,704	960,103	957,178
Southern Europe	332,769	50,007	693,595	439,295	3,260,825	582,720	3,112,718	1,593,817
Western Europe	3,840,848	1,107,365	1,103,989	542,957	6,024,633	741,588	5,488,083	2,974,052

This dataset can be interpreted in a number of different ways. However, to facilitate accessibility and avoid overwhelming complexity of story or visualisation, a first cut of the data can be generated which can summarise the state of the world for a given year and summarise the extent of migration to each country/region chosen. Given the different geographical levels range from continent, region down to country we chose to use a 'sunburst' plot to adequately represent the hierarchies, allow the user to drill-down and explore the data as well as, gauge the relative volumes across each unit.

Following an initial familiarisation with the dataset, the *flow* between areas can be explored utilising a 'circular plot' or a type of rounded Sankey diagram to outline the flow from a region or country to another.

The preparation and utilisation of this dataset with the respective visualisation required careful transformation. To maintain a geographical focus in our analysis to Europe and Asia, all other areas are filtered and arrange in the same order which the raw dataset does not necessarily adhere to. Subsequently, python is utilised to turn the each excel row into a one-dimensional matrix and finally transfer all generated matrices to our HTML document in order to visualise.

3.2 | National migration data – Tencent Data Store – internal China mobility

In contrast to our previous dataset, the data on internal Chinese mobility approaches more of a 'big data' dataset, purely from a volume dimension, however it lacks the broadly expected fuller suite of characteristics as per some recent debates in the literature (De Mauro et al., 2016; Ward and Barker, 2013). Split into two parts, covering population inflow and outflow with circa two million records in each covering a period of 18 months at a daily temporal resolution from January 2018 to June 2019.

In addition to origin and destination denoted in Mandarin, a percentage split is included between three modes of transport: car, train and flight. These percentage are calculated on the basis of the principal measure referred to as a 'hot degree' (referred to as intensity in this report) the values of which range from single digit thousands to millions depending on the city (Heat.QQ, 2020).[1]

Figure 4 | Hot degree formula

$$\begin{aligned}\textbf{\textit{hot degree}} &= weight \times passanger_{count} \times Air_{ratio}\\ &\quad + weight \times traveller_{count} \times Train_{ratio}\\ &\quad + weight \times trips_{count} \times Car_{ratio}\end{aligned}$$

Notably missing from this dataset is relevant geospatial information such as coordinates, or membership of organisational units such as provinces or regions making it impossible to visualise points or draw connections between points. This proved to be a key data processing challenge that was overcome in two steps. Firstly, a generation of a reference dataset manually in order for the dataset to be explored meaningfully and proof of concepts created. Secondly, a matching of the unique set of Chinese character city names (circa 500) through the Google places API which returned coordinates but also Latin-character names. Addressing this challenge allowed us to proceed with the visualisation in the following way.

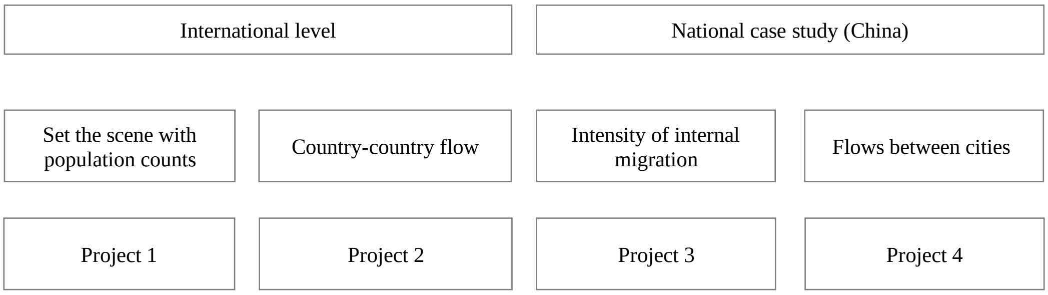
In line with our approach at the international level, so at the national level a two-step visualisation journey is proposed. Firstly, a view of the data to showcase the disparities in intensity of migration across China, allowing for a spatial dimension and allow for this to be explored across time, the temporal dimension. Second with a grasp of spatial-temporal patterns, we can incorporate the origin-destination element in a separate view for simplicity and showcase the change in connection among cities from one time period to another, as well as their relative intensity. This is crucial in informing the user not merely where people move from and to but allow the volume (or intensity) of people to be approximated at a city level.

[1] More information available in the appendix.

3.3 | Next steps and methodology

On the basis on the data presented in Sections 3.1 and 3.2, our proposed approach is to produce four distinct projects, two per dataset. The first of each pair, sets the scene for the user whereas the second enables the user to carry out more complex exploration. Figure 5 outlines the basis of project organisation.

Figure 5 | Modular approach representation



International level		National case study (China)	
Set the scene with population counts	Country-country flow	Intensity of internal migration	Flows between cities
Project 1	Project 2	Project 3	Project 4

It is important to note that there is no interactivity between the international set of analyses and the national Chinese case study. International level projects (1 and 2) address research question 1, while the remainder address research question 2 and its constituent parts.

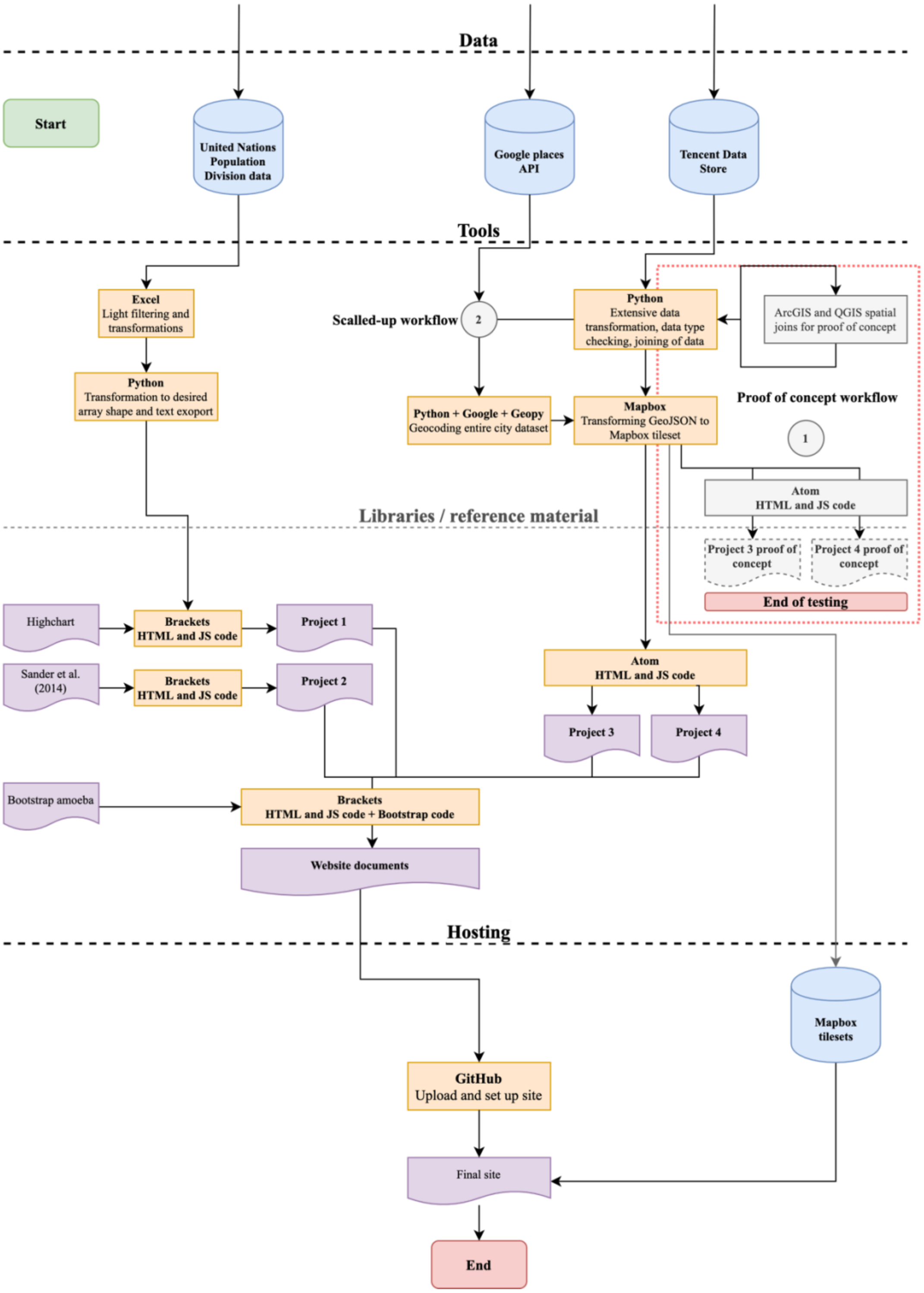
Section 4 below, outlines the tools and process undertaken to construct the final output with Section 5 breaking each project down into detail with specific design and technical considerations/decisions made.

4.0 | Technology stack

Summary of components brought together to make our output possible

From a technical perspective, a variety of data formats, tools, libraries and processing were used to produce the end website. This section breaks-down the key elements and serves as a means of transparency and reproducibility with Figure 6 acting as a visual representation.

Figure 6 | Data, tools, libraries, hosting relationships visualised


Data
Start
United Nations Population Division data
Google places API
Tencent Data Store
Tools
Excel
Light filtering and transformations
Scalled-up workflow
2
Python
Extensive data transformation, data type checking, joining of data
ArcGIS and QGIS spatial joins for proof of concept
Python
Transformation to desired array shape and text exoport
Python + Google + Geopy
Geocoding entire city dataset
Mapbox
Transforming GeoJSON to Mapbox tileset
Proof of concept workflow
1
Atom
HTML and JS code
Libraries / reference material
Project 3 proof of concept
Project 4 proof of concept
End of testing
Highchart
Brackets
HTML and JS code
Project 1
Sander et al. (2014)
Brackets
HTML and JS code
Project 2
Atom
HTML and JS code
Project 3
Project 4
Bootstrap amoeba
Brackets
HTML and JS code + Bootstrap code
Website documents
Hosting
Mapbox tilesets
GitHub
Upload and set up site
Final site
End

4.1 | Data sources and formats

As section 3 covers, we built on analysis and site around data from the United Nations and Tencent for international and national projects respectively. However, Table 3 outlines additional data sources and formats that had to be integrated to make data usable.

Table 3 | Data sources and formats

Data source	Format	Additional datasets to enhance/clean
United nations (International migration statistics)	Excel	Filtered in Excel and processed in **Python** and transposed into correct **array** form usable with Project 1 and 2 and embedded directly into the visualisation code.
Tencent (national migration intensity)	Comma separated file	Raw **.csv** files processed through **Python** as Panda **data frames**, correcting for **encoding errors** of Chinese city name characters. Further joining occurred between the data frames, **shape files** (.shp), geocoded **JSON** returns from **Google's geocoding API** before finally exported as **GeoJSON** files that were submitted to the Mapbox service for **tileset** conversion

The United Nations dataset, being curated proved to be clean and easy to use. The principal challenge lay in understanding the correct array shape and format to be ingested by the visualisations created.

The Tencent data in contrast, lacked key spatial information (such as coordinates) and Latin character names. Firstly, we added this data with a manually generated shape file in order to not have a bottleneck in the design of our visualisation before proceeding to make geocoding API calls to geolocate the entire dataset and pull Latin-character names returned in JSON format and merged into the master data frame in python. Finally, the choice was made to go from GeoJSON generated by python to Mapbox's own tileset for to externalise hosting and leverage the efficiency provided by this format.

4.2 | Data processing tools

As outlined in the previous section, Python and Excel were the primary tools used to interact with the bulk of our data, more detail on those and additional tools provided in Table 4.

Table 4 | Summary of data processing tools

Tool	Primary use	Description and rationale
Excel	Project 1 & 2 – critical	Highly flexible and accessible. Was well suited for UN data manipulation and one-off amendments to our raw data (such as temporal filtering) that did not have to be revisited.
Python	Project 2,3,4 – critical	Versatile and popular programming language with a rich array of packages that made it a suitable candidate for systematic understanding and processing/enhancing of our data. We utilised Python 3 with Jupyter notebooks to facilitate exchange of workflows between team members and allow for mutual review of code.
ArcGIS	Project 3,4 – exploratory	Both tools were used for equivalent tasks but across Windows and Mac OS platforms. ArcGIS added coordinate data to a sample of key cities in order to allow web development of both Project 3 and 4 to continue. Furthermore, spatial joins and connecting lines were drawn to illustrate the proof of concept for Project 4. Easy and accessibility for manual intervention is a notable feature of these tools in contrast to Python which is better suited for bulk tasks. With success in our trials, we then translated our manual processes to Python script and automated them.
QGIS	Project 3,4 – exploratory	
Mapbox	Project 3,4 - critical	Utilised to convert GeoJSON to tileset format and host datasets – no further data intervention

***Note**: working files available as part of project in GitHub repository*

For our final output our priority was to utilise widely accessible / cross-platform tools as much as possible. However, the use of ArcGIS and QGIS (its open source counterpart) were critical in utilising the different skills of the team to generate working proof of concepts that were later turned into scalable and automated workflows in Python. This enables us to share a more replicable piece of work, but crucially, greatly enhance future updates for this project.

4.3 | Text editors, libraries and reference code

No strict requirements were applied to selection of text editors except for individual proficiency. We utilised Atom and Brackets to write our HTML and JavaScript code (Atom.io, 2020; Brackets.io and Adobe, 2020).

Central to this section are the external libraries and code / best practice references utilised to varying degrees across each of the four projects.

Table 5 | Summary of reference code and libraries

Library	Use in project	Description of use
High charts	Project 1 – use of library	Open Library with a number of starter templates for different visualisations. As the purpose of this project is to set the scene emphasised was placed on simplicity and effectiveness. This enabled for experimentation and choice of 'sunburst' chart to visualise data that was transformed into the required format (Highcharts, 2020).
Sander et al. work (2014)	Project 2 – adaptation of existing project code	Project 2 was extending the original work beyond its original timeframe. As a result, the existing project resources were leveraged from their open GitHub repository in order to maintain consistency of presentation. Care was taken to review the licence under which the work is shared which allows for adaptation provided appropriate recognition is made (null2 GmbH Berlin, 2015).
Mapbox.gl	Project 3,4 – use of library	Both JavaScript and CSS components used in visualising and interacting with the mapping layer created to allow actions such as: double click to zoom; pop-up message generation with additional information for a point on the map, and click to show/hide layers.
Google Fonts	Project 3,4 – use of library	For greater consistency with our main page, the default fonts used from Google were used.
Bootstrap	Website framework to house above projects	To house the above projects, a bootstrap website template was selected as it includes features such as cross-browser and device compatibility. However, care was taken to adapt this starting point to accommodate the requirements of this work (BootstrapMade, 2020).

***Note**: code available for review in GitHub repository*

Focused on adaptation and enhancement of existing work projects 1 and 2 were able to effectively build on existing assets to visualise the data and present findings. Projects 3 and 4, given their bespoke requirements built upon the Mapbox.gl libraries to construct views from the ground-up. All were housed within the customised 'Amoeba' reference template that was tailored to this specific use case.

4.4 | Data and website hosting

Finally, for a project of this kind, we continued to focus on open source solutions. As outlined before, Mapbox was used to host all key data components for Projects 3 and 4 whereas data for Projects 1 and 2, along with the website files itself were hosted on GitHub as part of the repository for this project. This provides a one-stop-shop for aspects of the project that are open while allowing us to show, without sharing the Tencent data which is not open source itself.

With this background, the remainder of this report focuses on three aspects at a project level. Firstly, technical (code or visualisation) considerations together with planning ahead of the projects. Secondly, implementation challenges and how they were overcome. Thirdly, findings and possibilities for the user to explore the respective datasets.

5.0 | Technical and visual considerations

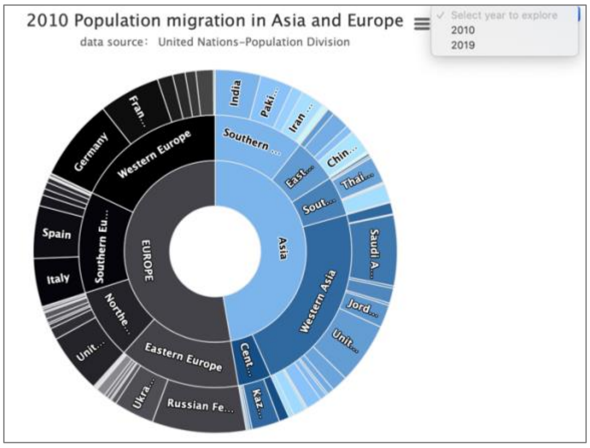
Outlining technical and design considerations and challenge for each project

This section focuses mostly on the completion of the two pairs (total of four) visualisation projects, our respective considerations in the process of reconciling simplicity and accessibility of each visualisation together addressing our core research questions. Of note are the distinct visual identities between each pair, utilised for a starker differentiation of the geographical level and dataset that is utilised with each. Finally, additional project specific challenges will be expanded on.

5.1 | Project 1 | Count of population departing a country

This sunburst diagram shows the population migration in Asia and Europe in either 2010 or 2019. The reason why choosing sunburst diagram is that the pie chart can display the proportion of migration intuitively and users can have a good interactive experience from it. As shown in the diagram, the pie contains two parts: the grey/black portion represents Europe and blue side represents Asia.

Figure 7 | Project 1 screenshot



The reason for choosing this approach is that sunburst fits the principle of data visualization: displaying complex data with simple image graphics, large amount of information, beautiful and easy to read. The proportion of each level in the hierarchy is represented by a circle. The closer to the origin, the higher the level of the ring. The innermost circle represents the top layer of the hierarchy, and then the reader can view the proportion of the data layer by layer. It looks like a doughnut chart or a pie chart, but it is actually much more complicated-because each sector of sunburst can continue to be divided down. This has become a multi-level, multi-dimensional distribution visualization method. In this visualization, it can not only track the amount of people, but also show the proportion of various countries, and also automatically summarize the status of different levels, so that users can truly understand the specific composition of the data.

Although not intended for comparison between points in time, functionality was added to explore multiple years so that (given the temporal dimension of Project 2), a user could familiarise themselves with the requisite data ahead of time.

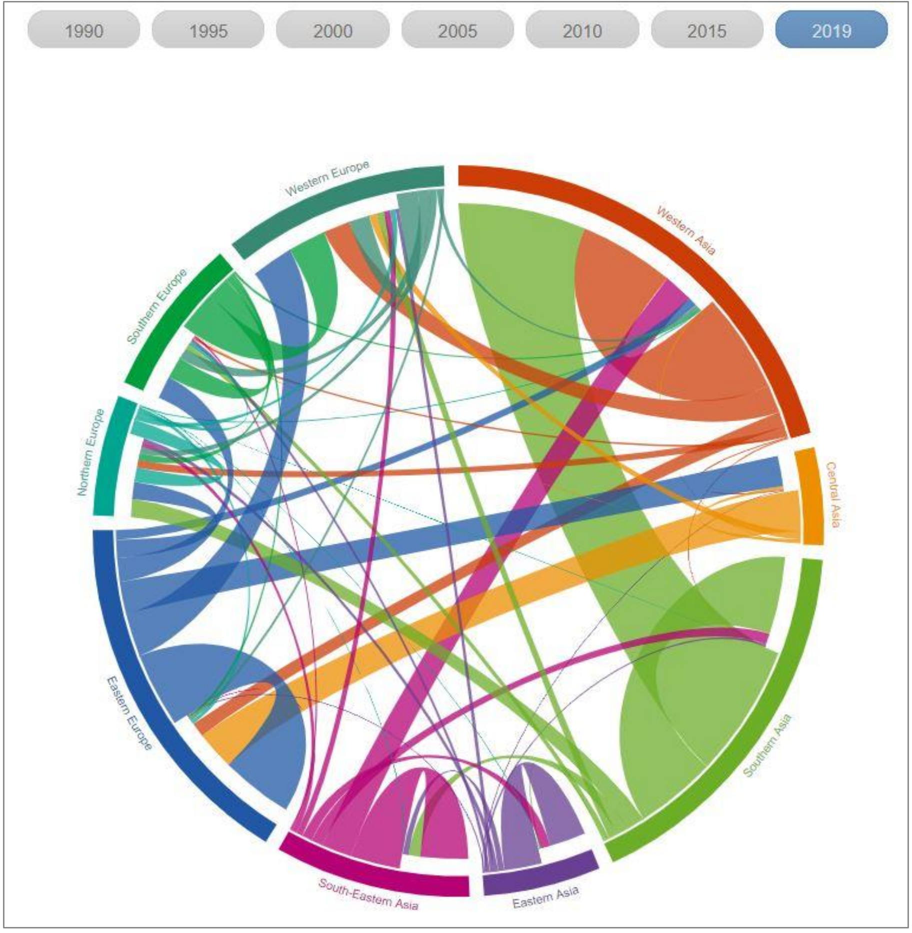
When clicking Asia, a new sunburst diagram appears, showing the specific population migration in different parts in Asia. We can continue click on one part such as Western Asia, then the pie chart including all countries in western Asia showed up, where the area of each country represents the population migration. The number of migrants appears when the mouse move to the country or the continent.

Although it is easy for users to view the required data by clicking in the sunburst diagram, this diagram is unable to show the concrete flow from one country to another. We can only have a rough understanding of the general proportion of each country and each part of continent. Therefore, it is necessary to show the detail of the flow between countries and continents by applying different diagrams.

5.2 | Project 2 | International migration flows | updating Sander et. Al (2014)

Building on project 1, this research continues to visualize data in detail about Asia and Europe. Projects 2 aims to show the relationship between the migration population. Inspired by 'The Global Flow of People' project created by Nikola Sander, Guy J. Abel & Ramon Bauer, this part designed a circular Sankey-style diagram about Asia and Europe.

Figure 8 | Project 2 screenshot



The reason for choosing Sankey is that this type of graph is extremely suitable for representing flow. The width of the extended branches in the figure corresponds to the size of the data traffic. The sum of the widths of all main branches should be equal to the total width of all branches. This kind of graph maintains the balance of inflow and outflow energy and is very suitable for visual analysis of flow data. In the process of analysis, the direction of the line is also the direction of the data flow. We could analyse the change of the direction of the data flow according to the direction of the line. At the same time, we also need to pay attention to the change of the line width, which presents the amount of the data.

Therefore, we can clearly see how many populations flow out from a country or region in a certain time period. The thickness degree of each line represents the number of populations movement. It is simple to recognize how is this country or region behaving.

The dataset which is on world migrant stock processed by deleting some columns and rows. It only needs to be kept the data related to Asia and Europe between 1990 and 2019. In this cleaned dataset, Asia is divided into western Asia, central Asia, southern Asia, eastern Asia and south-eastern Asia. Europe is split to eastern Europe, Northern Europe, Southern Europe and Western Europe.

Brackets is used to build HTML environment. In this webpage, a year filter is added on the top. Years are separate from 1990, 1996, 2001, 2006, 2011 and 2016 to 2019.

In this Sankey diagram, we can clearly see close ties between Europe and Asia as illustrated by the flows of people. The most population movement in western Asia, southern Asia and Eastern Europe. There is no populous country in them. Therefore, it might cause by the people need jobs abroad.

Regarding our overarching research question for the international level, we can observe certain shifts such as the gradual increase of west Asian migration and the general proliferation of inter-regional as opposed to previously intra-regional migration. Therefore, we do not observe a continuation of the trend represented in the original work despite the use of a less precise dataset.

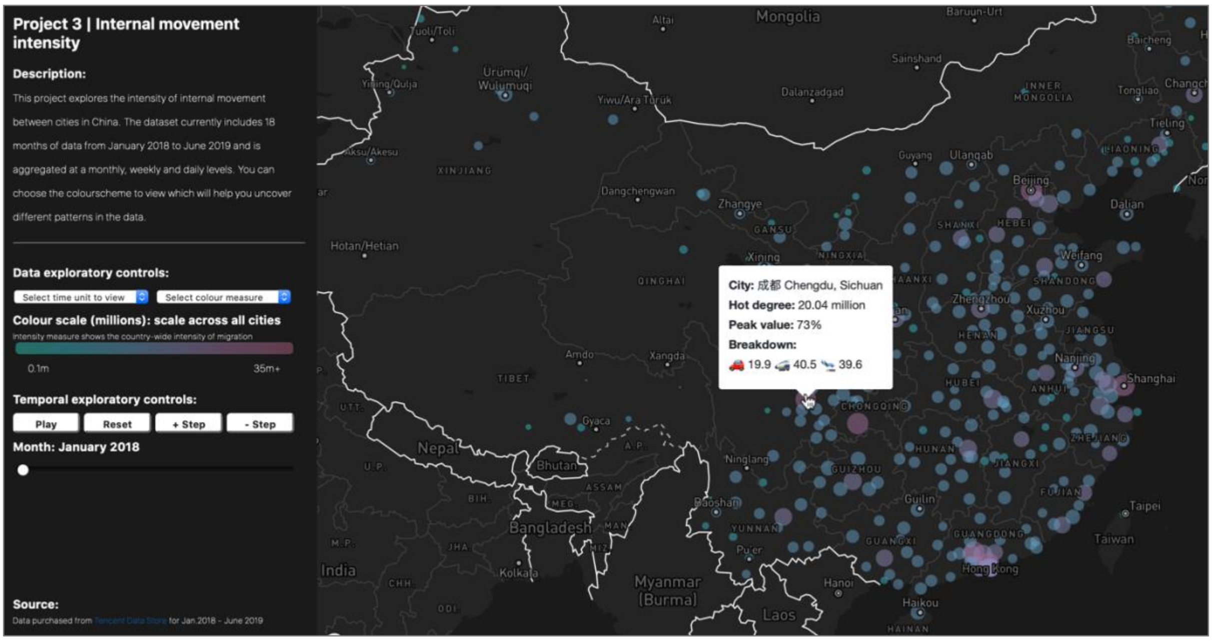
An observation of note, that ties to our next pair of projects is the fact that China, as the world’s most populous country, has low levels of external migration and therefore piques our interest in understanding its internal movements which we expect to strongly contrast with its connected but relatively low profile.

5.3 | Project 3 | Intensity of internal migration in China

In begging our exploration of China, a similar approach is taken to our international work in setting the scene with Project 3 and diving deeper into origin-destination flows in project 4.

The purpose of project 3 is to allow the user to explore the spatial-temporal element of our China dataset and explore patterns pertaining to the relative intensity of migration between cities (our intensity scale) or within the city's temporal history (our peak scale). Figure 9 shows a preview of the visualisation.

Figure 9 | Project 3 screenshot



Project 3 | Internal movement intensity

Description:

This project explores the intensity of internal movement between cities in China. The dataset currently includes 18 months of data from January 2018 to June 2019 and is aggregated at a monthly, weekly and daily levels. You can choose the colourscheme to view which will help you uncover different patterns in the data.

Data exploratory controls:

Select time unit to view | Select colour measure

Colour scale (millions): scale across all cities

intensity measure shows the country-wide intensity of migration

0.1m 35m+

Temporal exploratory controls:

Play | Reset | + Step | - Step

Month: January 2018

Source:

Data purchased from Tencent Data Store for Jan.2018 - June 2019

City: 成都 Chengdu, Sichuan

Hot degree: 20.04 million

Peak value: 73%

Breakdown:

🚗 19.9 🚄 40.5 🛬 39.6

The overall design principals are, as mentioned previously such that continuity is established between projects 3 and 4, setting them visually apart from our international projects. Furthermore, given the greater volume of data, the layout takes into account the need for finer user interaction controls and aims to balance that with simplicity. Consequently, an approach that splits the view into a left third and right two-third is used. This not only takes allows a new viewer to be naturally guided through the descriptive content at the top left and gradually read down to understand the means through which data can be adjusted, it provides sufficient screen real estate to view the data in its entirety.

Further design considerations revolve around three crucial components, our base map, and representation of cities which comprises of point-size and point-colour. The latter two aspects in particular elaborate on the interplay of the control panel in the left third and the principal visualisation in the right two thirds.

5.3.1 | Base map | High contrast for visibility

For our base map, we utilised a high contrast dark theme. Firstly, it in itself contrasts with the white-themed international level analysis but secondly, allows for variety of colours to be interpreted more easily. The colour scheme of the base-layer is translated into the colours of the left-third section for continuity, emphasizing the grey background (which matches the water of the map) and light text.

5.3.2 | Point-layer | Size as ‘hot degree’ (intensity)

In order to understand the challenges and design decisions regarding the size of each point, we must turn back to our dataset and the values and distribution of our ‘hot degree’ measure as illustrated by Figure 10. Parts A and B show the data in its raw form and transformed using a log function respectively. Every ‘tick’ along the x-axis represents an individual city and the points along the y-axis represent the values of intensity, giving us an understanding of the range involved (larger versions available in the appendix).

Figure 10 | Visualising all monthly points per city (ranked by mean heat)

Figure 10A | Raw ‘hot degree’ value by city

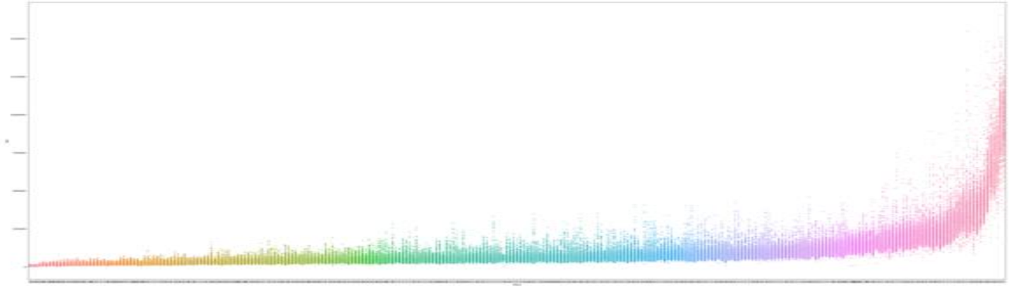
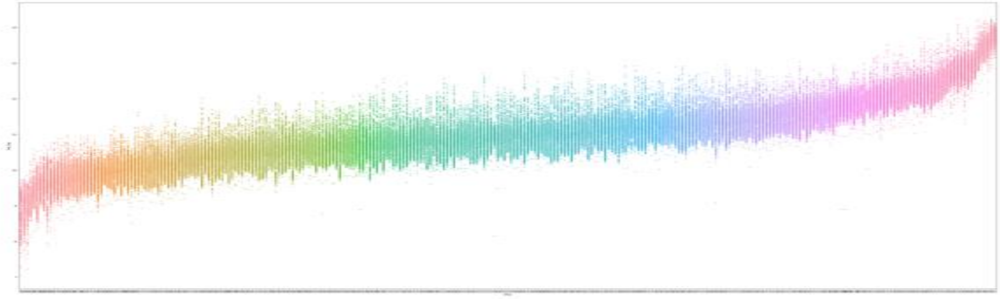


Figure 10B | Logged ‘hot degree’ by city



The trade-off is between accurate reflection of the values and noticeable enough changes in the visualisation itself. As we can see, cities in the light purple towards the right end of the graph have consistently larger values and therefore require cut-off stops for colour ranges to be designed in such a way that allows a user to understand the specific peaks and troughs of each city.

For the size of the bubble the logged values were selected to fulfil this purpose and consistently applied across the three-temporal aggregation of day, week, month. This normalisation of data allowed for more evident changes in size across zoom levels without too many instances of overlapping points where the larger cities obstruct smaller surrounding points with data.

5.3.3 | Point-layer | Using colour as complimentary visual cue

Colour adds an additional dimension that had to be calculated.

Firstly, in contrast to size, colour stops were designed on the basis of actual ‘hot degree’ values as the constraint of obstructing the map or neighbouring points did not exist.

This therefore acts as a secondary cue to the user reinforcing the messages regarding the relative intensity between cities.

However, given the temporal nature of this dataset, an additional measure was calculated that focuses on the intensity of migration from the perspective of each city. Normalising all their values on a scale between 0 and 1. This is a powerful option to explore as it clearly shows peaks across the nation at key points in time, such as late January/early February where New Year celebrations prompt larger internal movements. Given the consistent nature of this measure a simple ranged colour scheme between blue and red with a yellow midpoint could be used consistently across all cities and temporal resolutions.

5.3.4 | Challenges along the way and additional features

We have previously discussed the challenges surrounding geocoding the dataset. Additional challenges had to be overcome when it came to visualising the volume of data and providing the right metadata to the user to further their understanding.

Originally at a daily temporal resolution, we reached bottlenecks with the format of our GeoJSON and the lack of a relational dataset that would avoid the duplicative nature of the data we were uploading to Mapbox. In effect, every data point, for every day became a separate point, despite the fact only one would be shown at a time. This imposed performance constraints, as well as visualisation constraints, such that points for individual days are not visible at lower zoom levels. Our implemented solution addresses this in two parts.

Firstly, aggregate data from its raw daily form to weekly and monthly datasets. This not only greatly reduced the size of our data, it also provided easier to understand temporal series that could be viewed in a manageable amount of time. Furthermore, as jumps between points in time were achieved through filtering of our dataset in tileset form, the size of the dataset has an impact on the speed of the filtering process. Therefore, we were able to have a higher refresh rate at 500 milliseconds as opposed our previous 1 frame a second. Secondly, for those wishing to explore the daily level data, we implement visual prompts to zoom further into the layer along with functionality whereby click on the map takes them to their area at the desired zoom level.

Finally, in order to showcase the additional data we had such as the breakdown between means of transport, we implemented map pop-up that displayed critical information when a user hovers over a point.

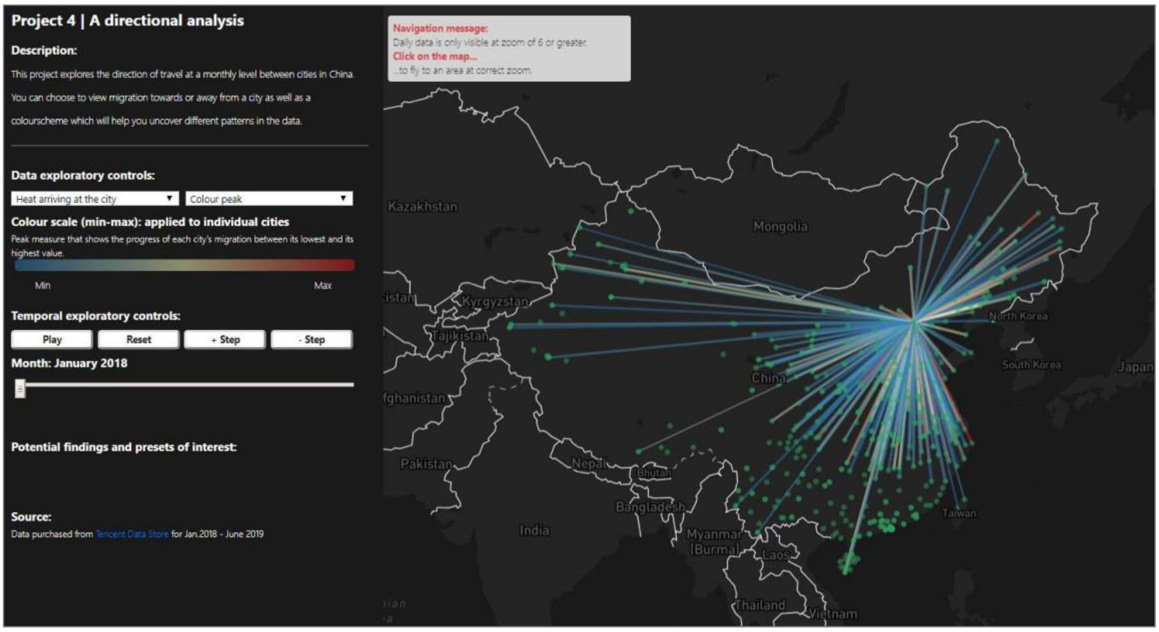
5.3.5 | Findings

The user can explore the data in a number of ways. However, pertaining to our research questions. We can indeed observe seasonal patterns with the time around February showing particularly high peaks and quiet periods thus addressing questions 2.2. We can partly address the intensity point of question 2.1 in that we see, as expected, the higher intensity along the coastal periphery. However, to analyse the directionality, we turn to Project 4.

5.4 | Project 4 | Population flow in China

Project 4 is based on the visualization project 3, and further explores the direction of travel. This not only allows users to understand the intensity of the migration flow between cities, but also realize people's origin, destination and intensity of travel across given paths.

Figure 11 | Project 4 screenshot



Project 4 | A directional analysis

Description:

This project explores the direction of travel at a monthly level between cities in China. You can choose to view migration towards or away from a city as well as a colourscheme which will help you uncover different patterns in the data.

Data exploratory controls:

Heat arriving at the city ▼ Colour peak ▼

Colour scale (min-max): applied to individual cities

Peak measure that shows the progress of each city's migration between its lowest and its highest value.

Temporal exploratory controls:

Month: January 2018

Potential findings and presets of interest:

Source:

Data purchased from Tencent Data Store for Jan.2018 - June 2019

Navigation message:

Daily data is only visible at zoom of 6 or greater.

Click on the map...

...to fly to an area at correct zoom.

At the outset, we must state that as these two projects work closely together, the same visual language, layout and rules were applied to many aspects such as page organisation, base layer and colour pallet. For brevity these points will not be elaborated in great depth to avoid repetition, however the reader may refer to section 5.3 for more detail.

The biggest advantage of using flow is that it can give us a direct visual impression; what is more, it can help user know the city's influence in their province or area. As with project 3, the reach and intensity of our 'hot degree' measure varies greatly between urban centres with centres such as Shanghai and Beijing understandably dominating as the financial and political capitals respectively.

Line thickness is utilised using the same logged measure as is used in Project 3 for circle dimension. However, given the high frequency for certain cities the variation in thickness is faint at lower zoom levels. To counter this phenomenon and in keeping with the structure of this pair of projects, colour is utilised to reinforce this datapoint. Therefore, we are able to avoid line occlusion and convey the principal message.

Furthermore, as we are visualising a number of lines per one point in time, our Python dataframe size and subsequently GeoJSON files increased rapidly in size. As a result, they proved impractical to visualise in Mapbox in their current state. Consequently, only data at monthly level is used. Nevertheless, in order to avoid the oversaturation of the map that would force Mapbox to limit our minimal zoom level, each month is uploaded as a separate layer. Therefore, in contrast to Project 3, the time slider does not filter our tileset but instead only shows the relevant layer based on the layer IDs assigned on load. Finally, a novel addition compared to the previous project is the ability to see top destination points as well as all origin points.

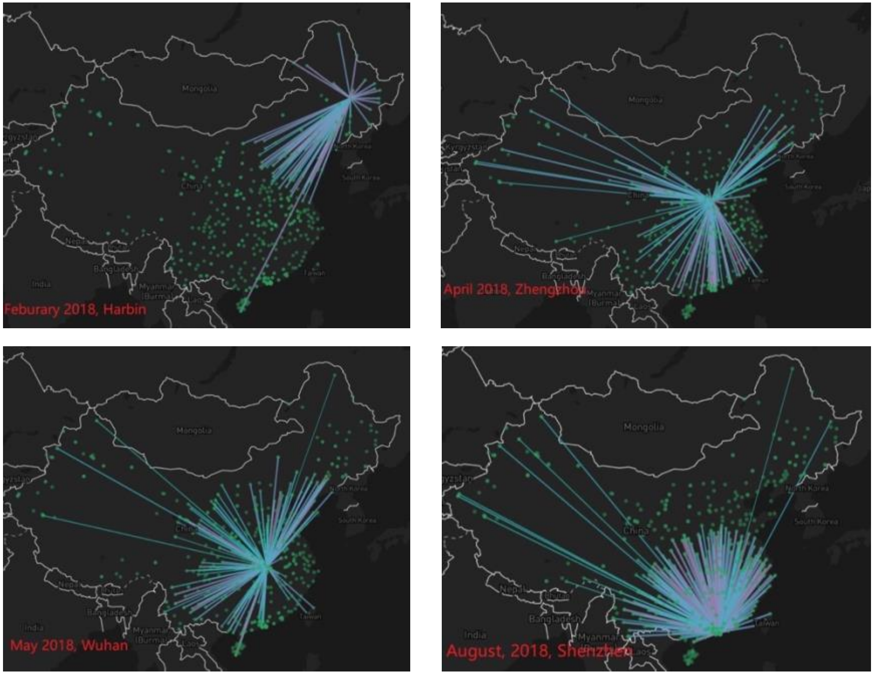
5.4.1 | Findings of interest

A key finding of project 3 is that the cities with a higher ‘hot degree’ tend to be the 32 province capitals of China and other centres under the jurisdiction of central government. This is corroborated by findings in project 4 with the reach and number of connections per city being extensive and often covering up to half of China’s land area. Furthermore, we observe this phenomenon holding throughout the seasonal cycles.

Some explanatory factors will be, as per our literature review, the economic power of these cities that attract individuals through improved employment and standard of living prospects. Furthermore, these cities act as key nodes in China’s internal travel network and thus will inevitably serve as points of transition as well as end destinations. This is something which we cannot necessarily separate in this dataset.

Chongqing is an interesting case study and point of discovery in this dataset in that on occasion, certain points although defined as a single city are in fact a collection of cities which helps explain the high level of ‘hot degree.’ This explanation can be further supplemented with historical and political factors that make this city a particularly busy thoroughfare. These are the kinds of points we do not contextualise but hope to prompt our user to want to go an explore.

Figure 12 | City link examples



There is also a very interesting phenomenon, we can find that the radiation range of each city will be affected by distance, such as Harbin and Shenzhen (see figure above). These cities have a strong population mobility relationship with the cities around them. As the distance increases, the relationship becomes weaker. But if the city rank is higher, this phenomenon will be avoided to a certain extent, such as Beijing and Shanghai. But it should be noted that Wuhan and Zhengzhou. These two extremely important transportation hub cities in China do not have a very strong population mobility relationship with Shanghai, Zhejiang and Jiangsu Province. This may be speaking to the phenomena in the literature on an increasing prevalence in intra-province migration as development has become more uniform across the country as goes some way to address the directionality component of question 2.1.

6.0 | Conclusion

Summary of our work, addressing research questions and highlighting further improvements

In conclusion, this report outlines the academic drivers, the data, tools and projects that make up our ‘On the move’ project and final website.

Four pieces of visualisation work are carried in pairs showcasing trends at an international level by building on existing repositories/codes and effective practices and at a national level in China where we tested the use of a novel dataset with bespoke visualisation approaches. Our focus was, to the extent possible on simplicity and interactivity that allow our users to explore the data at hand. The below summarises our two pairs of projects.

Project 1 used the sunburst diagram to visualize the proportion of migrants in various countries in Asia and Europe in either 2010 or 2019. Users can have an intuitive interactive experience and drill down into the data, as well as switch between the years they wish to explore. **Project 2** made a Circular Sankey diagram to show the detailed outflows of migration between Asia and Europe. This utilised and built on the work by Sander et al. (2014) and their code that was published alongside their research. Our extension focused on the addition of an alternative data source across a longer time horizon in order to understand any changes since its original publication.

Project 3 applied the bubble and intensity map to explore intensity of population migration in China across different temporal resolutions. Users are able to explore the spatial-temporal element of our China data set and compare the relative intensity of migration between cities and across time for a given city. **Project 4** as an extension on project 3, used a origin-destination map to discover the origins and destinations of movement between core movement hubs in China. Both the intensity of the migration flow between cities and people’s departure and destination can be found on the map. Users can get a direct visual expression by observing the flow between cities and then find which city has the largest radiation range.

We find that our work allows to argue that population migration trends are in line with their historical trajectories that see gradual increasing movements from intra-regional to inter-regional travel and thus addresses our first research question. Secondly, we find that the dynamics of the mobile phone generated movement data, does broadly align with the internal movement dynamics of China we would expect based on datasets from national authorities. This alignment in direction, intensity and spatial phenomena that show the prevalence of intra-province travel suggest that there may be a use-case for this dataset as a more readily available proxy for broader systemic phenomena that are otherwise assessed using dataset that are updated much less frequently.

Room for improvement remains. Further work can focus on a more unified user experience through a single scrolling site, as well as, additional contextual visualisation that compare temporal trends at the international level and perhaps add more contextual information for each city in our national China-centric work. Finally, however, the challenge of effectively processing large volumes of data will be the main hurdle to overcome in order to most effectively and efficiently present data at a national level.

7.0 | Bibliography

Abel, G.J., Sander, N., 2014. Quantifying global international migration flows. Science 343, 1520–1522.

Allen, W., 2018. Representing freedom and force: how data visualisations convey the complex realities of migration.

Atom.io, 2020. Atom.

BootstrapMade, 2020. Bootstrapmade [WWW Document]. Free One Page Bootstrap Template – Amoeba. URL https://bootstrapmade.com/free-one-page-bootstrap-template-amoeba/ (accessed 5.15.20).

Brackets.io, Adobe, 2020. Brackets.

Chan, K.W., 2013. China: internal migration. Encycl. Glob. Hum. Migr.

Chan, K.W., 2001. Recent migration in China: Patterns, trends, and policies. Asian Perspect. 127–155.

Chan, K.W., Wang, M., 2008. Remapping China's regional inequalities, 1990-2006: A new assessment of de facto and de jure population data. Eurasian Geogr. Econ. 49, 21–55.

De Mauro, A., Greco, M., Grimaldi, M., 2016. A formal definition of Big Data based on its essential features. Libr. Rev.

Doyle, M.W., 2004. The challenge of worldwide migration. J. Int. Aff. 1–5.

Edmonston, B., Lee, S., 2018. Global migration and cities of the future. Can. Stud. Popul. Arch. 45, 33–42.

Goodkind, D., West, L.A., 2002. China's floating population: Definitions, data and recent findings. Urban Stud. 39, 2237–2250.

Hare, D., 1999. 'Push'versus 'pull'factors in migration outflows and returns: Determinants of migration status and spell duration among China's rural population. J. Dev. Stud. 35, 45–72.

Heat.QQ, 2020. Migration hot degree.

Highcharts, 2020. Latest stable code [WWW Document]. URL https://code.highcharts.com (accessed 4.27.20).

Liang, Z., Li, Z., Ma, Z., 2014. Changing patterns of the floating population in China, 2000–2010. Popul. Dev. Rev. 40, 695–716.

Ma, L., Tang, Y., 2020. Geography, trade, and internal migration in China. Cities China 115, 103181. https://doi.org/10.1016/j.jue.2019.06.004

null2 GmbH Berlin, 2015. Global migration data sheet 2013 [WWW Document]. URL https://github.com/null2/globalmigration (accessed 4.27.20).

Parsons, C.R., Skeldon, R., Walmsley, T.L., Winters, L.A., 2007. Quantifying international migration: A database of bilateral migrant stocks. The World Bank.

Poulain, M., 2008. European migration statistics: Definitions, data and challenges. Mapp. Linguist. Divers. Multicult. Contexts 43–66.

Sander, N., Abel, G.J., Bauer, R., Schmidt, J., 2014. Visualising migration flow data with circular plots. Vienna Institute of Demography Working Papers.

Tjaden, J., Auer, D., Laczko, F., 2019. Linking migration intentions with flows: Evidence and potential use. Int. Migr. 57, 36–57.

Vezzoli, S., Bonfiglio, A., De Haas, H., 2017. Global migration futures: Exploring the future of international migration with a scenario methodology.

Ward, J.S., Barker, A., 2013. Undefined by data: a survey of big data definitions. ArXiv Prepr. ArXiv13095821.

Zhang, H., 2017. Opportunity or new poverty trap: Rural-urban education disparity and internal migration in China. China Econ. Rev. 44, 112–124. https://doi.org/10.1016/j.chieco.2017.03.011

8.0 | Appendix

8.1 | A | Team member contribution checklist for Visualization Outputs

Component	Component Owner (s)	Additional Support
Presentation	All Group members	-
Idea dev. for presentation	All Group members	-
Idea dev. for visualization	Antonios and Xiang Zhou	Diqiu Yang, Xin Zhao
Lit. review	Antonios	-
Report review	Antonios	-
Data cleaning	Data cleaned by Project owners	-
Project 1	Antonios and Xiang Zhou	-
Project 2	Diqiu Yang, Xin Zhao	Antonios and Xiang Zhou
Project 3	Antonios and Xiang Zhou	-
Project 4	Antonios and Xiang Zhou	-
Website	Antonios and Xiang Zhou	-

8.2 | A | Team member contribution checklist for Report

Section	Diqui	Xin	Xiang	Antonios	Total
1.0 Introduction	415			200	615
2.0,2.1,2.2,2.3 Literature review				1100	1100
3.0 Data	150			150	300
3.1 UN		200		150	350
3.2 Tencent			200	150	350
4.0 Technology	200			400	600
5.0 Visualisation			100		100
5.1 Proj. 1		100	150		250
5.2 Proj. 2	350		150		500
5.3 Proj. 3			100	1000	1100
5.4 Proj. 4			750		700
6.0 Conclusion		420		100	400
Total	**1115**	**850**	**1300**	**3150**	**6365**

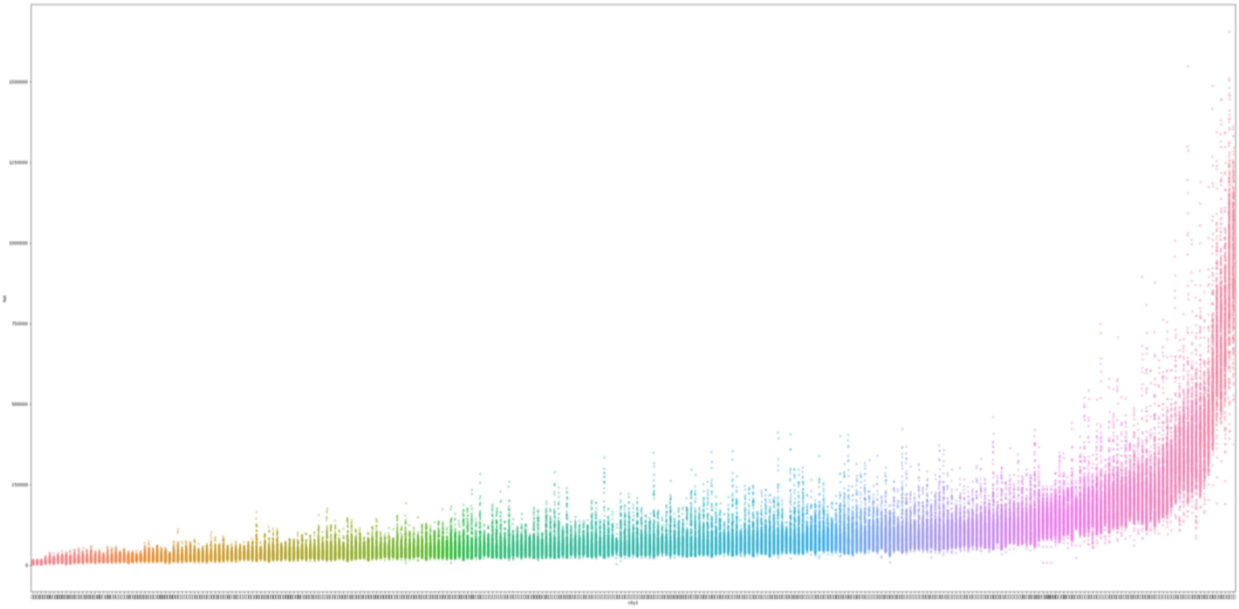
8.3 | A | Further hot degree description

The data is equivalent to that shown on the heat.qq platform (Heat.QQ, 2020). Our data is ata. higher level of accuracy this.

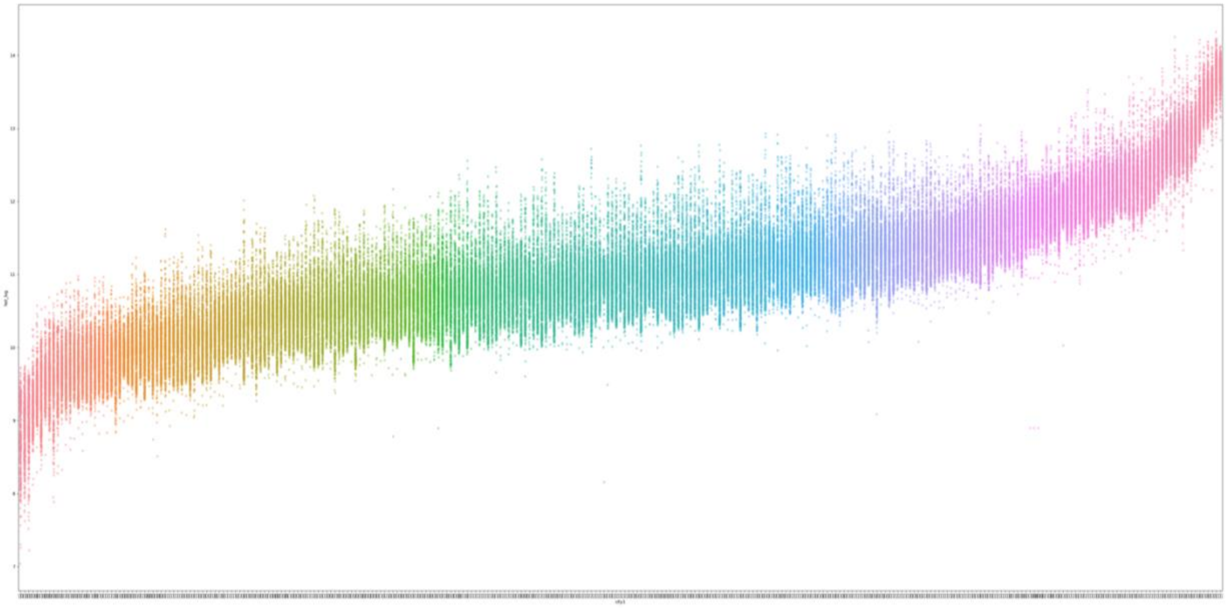
The hot degree after 2020 has been processed and turned into a percentage. Our data is before 2020, now it is out of print. This website is China's most authoritative data migration platform, and the specific calculation process of hot degree is the bottom line of this page.

8.4 | A | China data distribution plots

Part A | Real 'hot degree values' for figures aggregated to a monthly level.



Part B | Logged 'hot degree values' for figures aggregated to a monthly level.


hot_log