algorithms

MDPI

Article

Modalflow: Cross-Origin Flow Data Visualization for Urban Mobility

Ignacio Pérez-Messina [1,*,†], Eduardo Graells-Garrido [1,2], María Jesús Lobo [3] and Christophe Hurter [4]

1 Data Science Institute, Universidad del Desarrollo, 7610658 Las Condes, Chile; eduardo.graells@bsc.es
2 Barcelona Supercomputing Center (BSC), 08034 Barcelona, Spain
3 LASTIG, Univ Gustave Eiffel, ENSG, IGN, 94165 Saint-Mande, France; maria-jesus.lobo@ign.fr
4 ENAC, University of Toulouse, 31062 Toulouse, France; christophe.hurter@enac.fr
* Correspondence: igperezm@udd.cl
† Current address: Av. Plaza 680, 7610658 Las Condes, Chile.

Received: 14 October 2020; Accepted: 11 November 2020; Published: 15 November 2020

Abstract: Pervasive data have become a key source of information for mobility and transportation analyses. However, as a secondary source, it has a different methodological origin than travel survey data, usually relying on unsupervised algorithms, and so it requires to be assessed as a dataset. This assessment is challenging, because, in general, there is not a benchmark dataset or a ground truth scenario available, as travel surveys only represent a partial view of the phenomenon and suffer from their own biases. For this critical task, which involves urban planners and data scientists, we study the design space of the visualization of cross-origin, multivariate flow datasets. For this purpose, we introduce the Modalflow system, which incorporates and adapts different visualization techniques in a notebook-like setting, presenting novel visual encodings and interactions for flows with modal partition into scatterplots, flow maps, origin-destination matrices, and ternary plots. Using this system, we extract general insights on visual analysis of pervasive and survey data for urban mobility and assess a mobile phone network dataset for one metropolitan area.

Keywords: information visualization; flow data; urban mobility; mobile phone data; pervasive data

1. Introduction

Urban planning has been around since ancient times, but only at the beginning of the 20th century was it developed into an academic field, where it had its boom in the need of accommodating old and new cities to the needs of the industry. Thus, urban planning, along with transportation planning, are modern disciplines based on models developed in pre-digital industrial times. Today, urban planning is turning towards sustainable mobility, a new paradigm that changes how the relation of the city to the environment and people is understood. In this scenario, urbanism is faced with a major challenge: incorporating new data sources and data-driven methodologies into its framework. We aim at easing this transition by using visualization in order to give domain experts a contrasting view of traditional and new data sources against each other, also filling a gap between domains that has already been noted by the visualization community [1].

Owing to its complexity and massiveness, no data source alone can account for the whole phenomenon of mobility within a city. Different ways of measuring and approximating this phenomenon have been developed in order to understand and shape the city. Travel surveys have been a key tool, relying on a small, but carefully orchestrated, sample, access to census data and statistical craftsmanship. They are expensive and time consuming to produce [2] and they show heavy under-reporting [3]. This traditional source is contrasted with the information that can be extracted

from available and emerging pervasive data sources, such as mobile phone [4], Wi-Fi [5], and social network data [6]. This type of data are inexpensive and could potentially be used for real-time analysis, but, as a secondary data source for urban planning, it has its different methodological origin, which makes a comparison between datasets a challenge for both domain experts and data scientists. For example, as network data does not rely on self-report, trips, and mode of transport need to be inferred from waypoint traces, which is not an easy task, because of their relatively inaccurate temporal and spatial sampling, which introduces an uncertainty that was not present in this form in survey data (or at least, it does not have the same origin). We call datasets stemming from different methodologies "cross-origin".

Especially at the pilot stage, where a new dataset is being evaluated for its possible adoption and use, urbanists need to qualitatively assess the new data source. From what we have collected in interviews with experts, there are no protocols for this kind of assessment in use yet. Because official information on mobility is scarce, the travel survey is the main benchmark to which compare a dataset, which suffers from its own structural biases, and it is updated in the lapse of years.

In this unprecedented scenario, we developed Modalflow as a system for visualizing and comparing cross-origin datasets of urban mobility. Its name is derived from the two concepts that define its data workflow: flows and modality (as in modal split). Using juxtaposed and linked views in a notebook-style layout, it incorporates and adapts different visualization techniques for flow data including scatterplots, origin-destination matrices, edge bundling, flow maps (Figure 1), and ternary plots. Modalflow allows users to look into the phenomena of transport and, more importantly, to compare different sources on these phenomena, identifying each source by its methodological features. The main contributions of our work are:

- A visualization system with coordinated views derived from a set of considerations about the nature of pervasive and traditional data sources,
- The distribution-aware selection tooltip: an enhancement for selection techniques.
- The sinusoidal flow encoding: a new encoding for bidirectional and multivariate flows as edges of a graph.
- A validation of our approach with a case study using real-world cross-origin datasets.

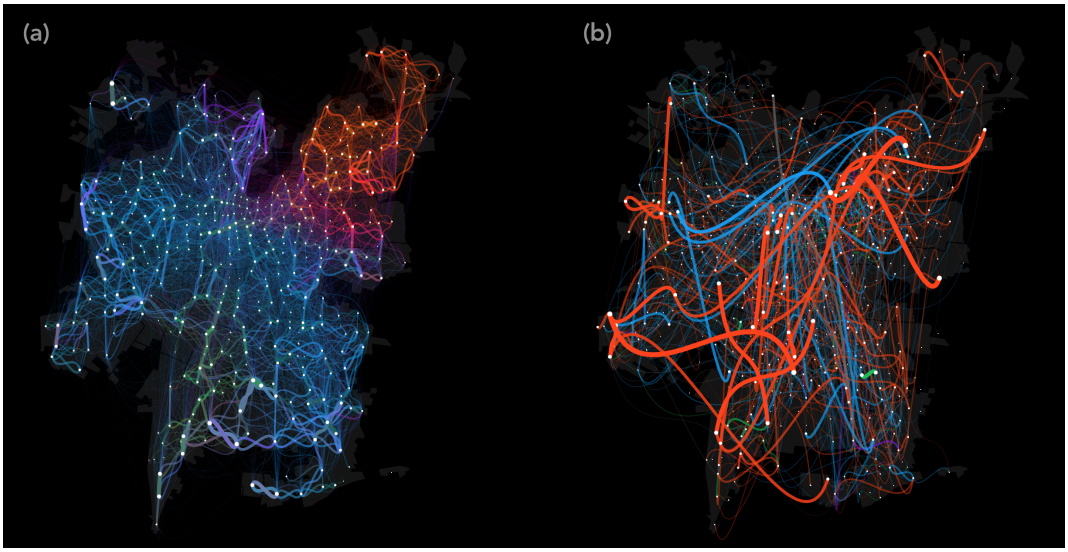

Figure 1. XDR (**a**) and survey (**b**) flows at night time.

We tested Modalflow while using two datasets from the Santiago metropolitan area: the last travel survey held in the city and a mobile phone network-inferred dataset, drawing insights on their differential qualities, strengths and flaws. We hope this system will foster critical evaluation of new, non-traditional data sources and their adoption by domain experts, and also improve the communication with the field of data scientists, which produces and makes use of these datasets.

2. Related Work

Mobility visualization has been mostly related to the visualization subfields of Infovis and Visual Analytics [7]. Here, we review the most common visualization techniques that are used for depicting flows in geospatial contexts and classify them according to the kind of data they are meant to be used with.

Origin-Destination Matrix. Origin-Destination (OD) matrices are a common representation used in transportation planning for low granularity data (such as municipal aggregation level). From a network visualization perspective, the OD matrix is a weighted adjacency matrix, where a continuous color scale is used in order to denote the flow magnitude between each node pair. Developments in this type of visualization have focused on improving readability and spatial mapping: MapTrix [8] enhances the traditional OD Matrix with interactive features and a linked geographical representation; OD Map [9] reorganizes the cells of the OD matrix to improve spatial pattern recognition; and, the OD Map can be combined with a tile map [10] to perform in more realistic scenarios. In spite of these improvements, OD matrices have not been used for visualizing networks of more than around a hundred nodes.

Graph. Graph-based flow visualization literature is greatly centered around reducing the hairball problem, i.e., visual clutter. We can divide them into those that operate some kind of simplification over the graph and then render the result with a linear technique [11–13], and those that make use of non-linear or iterative techniques for path calculation and rendering [14,15]. The techniques that were developed for this purpose rely on heavily distorting flow paths and edge appearance in order to improve readability. However, these developments have generally taken place without taking into account the particular characteristics of datasets and how they might behave on different data scenarios (e.g., do small changes in the data correspond to small changes in the result?) as it has not been studied how distribution can affect their performance at specific tasks (e.g., are scale-invariant as readable as random networks using this particular technique)?

Edge Bundling. Edge Bundling methods aim to visually simplify and aggregate trajectories, trail paths, or graph in dense visualizations. Recent surveys [16,17] provide a global overview of existing technique assets and usages. The first bundling algorithms were based on hierarchical data structure [18]. Recent techniques capture the statistical properties of the bundling [19] and enable combining multiple aggregation techniques [20].

Glyphs. Glyphs allow for encoding abstract flow information, confining it to a limited icon-sized space, and benefit from being visually independent from one another [21]. Following this line of research, Ma et al. [22] used the sunburst diagram as a model for a glyph, which encodes destination direction. In Andrienko et al. [23], radial diagrams are used to encode flow magnitude by direction and distance range. Pérez-Messina and Graells-Garrido expanded this idea to include mode of transportation [24].

Other Visual Techniques. Other kinds of transport mode-specific techniques have been developed for scenarios with constrained mobility, e.g., for metro use [25,26]. We are explicitly excluding heatmaps, which have been extensively used, as they are not able to retain network data.

Quantitative Data Source Comparison. There are studies reporting the quantitative differences between cross-origin sources: travel surveys and mobile phone data [27]/GPS data [28,29]. Additionally, survey inaccuracies from misreporting [30] suggest the need for a mixed approach [27,29].

This entails that, although flows, as data, may look like and be treated as the same matter when extracted from travel surveys or mobile phone network data, they are not. They are shaped by assumptions and hegemonic views that frame their methodological workflows [31,32]. This is a point that has been made by literature from critical visualization and digital humanities [33,34] and can be readily conceptualized by the data/capta paradigmatic opposition (data assumes information as "given", capta as "captured" and, thus, affected by our observation method) [35].

3. Design Requirements

Traditional data sources in urban planning, such as travel surveys, are grounded in self-reporting and explicit models, whereas non-traditional data sources, such as pervasive data, are grounded on unobtrusive capture and machine learning algorithms. This different epistemological status between data sources is the main concern of our design space exploration.

In our target domain, an established analytical workflow for evaluating and validating external data sources does not exist, thus visualization as a methodological approach plays a novel and central role in the solution of this problem. Here, we describe the domain problem, the considerations to approach this particular problem and the design requirements we extracted from them.

3.1. Domain Problem

In the target domain of application, there is not a standard methodology that can embrace current pervasive data sources, as these are mostly secondary sources, i.e., data that were not originally captured for the purpose of transport analysis. We consider pervasive data to be mode-unspecific and stemming from networks, such as social networks, Wi-Fi networks, and mobile phone networks. Primary sources, on the other hand, are datasets whose raw form and original purpose are inherently related to mobility (and not only as the consequence of an additional processing, as in XDR-inferred trips). There are other data sources that also enable mode-specific analysis, such as automatic passenger counting mechanisms in public transport [5], which, although highly granular, lose, to a certain extent, the origin and destination aspect of flows.

In the urban planning framework, there are three methodological approaches for generating primary data: qualitative, quantitative, and tracking. Qualitative methodology relies on focus groups and discourse analysis, whereas quantitative methodology relies on representative sampling, self report, and statistics. Tracking, on the other hand, relies on GPS devices that unobtrusively record an individual's path with a higher level of detail than any of the aforementioned and at a (relatively) lower cost.

Although pervasive data may seem to be more related to GPS data and tracking methodologies its coarser accuracy makes it more useful for quantitative analysis, such as reproducing travel surveys [36] or evaluating urban interventions of any kind [37]. The difficulties of working with network data are not just technical, as it is a massive data source, but also theoretical: there is an original methodological gap that separates mobile phone network data from survey data, which may, at least partially, explain its resistance to enter the traditional workflow. Recent developments show that public institutions are starting to make use of mobile phone data to understand mobility. In Spain, it is being used as a pre-census study [38] and, in Chile, it is being used to complement a travel survey in a small city in a pilot study [39]. This implies a rising need on tools that are aimed at experts to assess mobile phone data.

3.2. Considerations on Cross-Origin Dataset Comparison

Pervasive data as a secondary data source has its own qualities that sets it apart from primary source data. Here, we try to name and account for the aspects that make difficult the comparison between these sources, in the case of mobility flow data.

Trip closure. The definition of one trip does not share its origin in self-reported and predicted sources. For quantitative methodologies, what defines a trip as a unit is not part of any of these dimensions but a discursive element: the motivation of the sampled person (e.g., go to work, buy food, go for a walk, return home, etc.). Because tracking does not deal with persons but with devices, a trip is not a given entity, but it must be inferred from the data itself, usually from finding stay patterns that then become origin-destination features.

Representativeness. Surveys count with a relatively limited sample size, so sampling must be carefully orchestrated in order to achieve representativeness, the basic condition that guarantees its

usefulness as a tool for urban planning. The sample is based on the most recent census or projection of the census, which is the sampling universe, and then each point amplified based on expansion factors determined by statisticians. Because representativeness can only be achieved up to a certain level of granularity, observed patterns that are beyond that level should be considered to be an artifact (i.e., a value introduced artificially by the measuring tool). Mobile phone network data, although with a much higher level of sampling, may not count with this desirable quality.

Sampling density. The number of samples can vary by orders of magnitude between surveys and network data. Can sampling differences impair comparison? This enormous distribution difference can have an important effect when visualizing at finer granularity.

Dynamic behaviour. The temporal dimension is not equally accounted for by different sources. Travel surveys have an abstract temporality: a model is produced for a 'general day', which is then extended into laboral day and weekend behaviour for summer/winter seasons. It does not capture dynamic behaviour, as network data are able to.

Modality. From a transportation analysis perspective, modality cannot be overlooked. However, it is a hard dimension for pervasive data, as it usually has to be inferred. Probably due to this, visualization literature is usually focused on single mode (or absence of modality) analysis tasks.

3.3. Requirements

When considering the above synthesis, we extract design requirements that support a comparison of cross-origin flows given these aspects of conflict.

(R1) Comparing flows as multidimensional entities. Users should be able to visually compare different datasets in an origin-independent way, i.e., having the recorded trips aggregated into the common language of flows as multidimensional points, where different capture methods are comparable without hiding their inherent differences.

(R2) Comparing flows at different aggregation levels and ranges. Users should be able to filter and select flows according to spatial dimensions, origins and destinations, quantitative dimensions, time period, and visualize flows that are aggregated at different granularities.

(R3) Comparing mode split distributions. The system should allow comparing mode of transportation distribution in geographical space and along other dimensions.

(R4) Comparing flows in time. Users should be able to compare the effect of time (period of the day) on the mobility behaviour shown by different datasets and answer questions, such as "do different datasets show the same peaks during the day?" or "are daily modal cycles equally accounted for"?

4. Materials and Methods

In this section, we describe our test datasets and the visualization techniques developed for the system. An online demo is also available [40].

4.1. Data

Our cross-origin datasets comprise trips in Santiago, the capital of Chile, a city with almost 8 million inhabitants in 35 administrative units denoted municipalities within its urban area, further subdivided into 735 traffic analysis zones.

We define a flow as a unique origin-destination pair within a certain period (a partition of time), which has a magnitude (number of trips observed or predicted from origin to destination for the corresponding period) and a modal partition (the proportion of the magnitude corresponding to each mode of transport). Understood as a network where nodes are georeferenced objects, a flow is a weighted directed link with a modal partition uniquely defined by an origin-destination-period tuple. We also calculate a distance feature for each flow, as the linear geographical distance between origin and destination.

Travel Survey. The main source of travel demand and characteristics information for transportation purposes is the Santiago Travel Survey (known as "Encuesta Origen-Destino" in

Spanish, abbreviated EOD), which is held every ten years by the Chilean Transport Planning Secretary (SECTRA). In its last version from 2012, after surveying 100K inhabitants, the results showed that, in a typical working day, the transportation system has over 18 million trips, where walking is the most used mode of transport (34%), followed by public transport (26%), and private cars (26%) [41]. Arguably, most of the analysis of the travel survey focuses on the origin and destinations of trips, as well as their attributes, such as mode of transportation and travel distance. Given the sample size of the survey, only flows between municipalities are considered to be representative. Note that the traffic analysis zones of the city are defined by this survey.

Mobile Phone Network Data. As input data for Modalflow, here we work with a dataset of inferred commuting trips from mobile phone network data. Specifically, we use Extended Data Detail Records (XDR), which is a type of passive data that are used by telecommunication companies for billing Internet usage from mobile phones [4]. The dataset used in this paper was provided by the telecommunications company Telefónica Movistar, which had a market share of 30% in April 2017. The dataset contains trips during April 2017 for approximately 600K devices, which were extracted while using a previously published method [42]. The trip distribution inferred from this dataset has been validated with experts and through comparison with a travel survey, which implied that the market share covered by the dataset is enough to measure mobility patterns in Santiago [39].

These trips include a prediction of the several available modes in the Santiago transportation network, including bus, subway (metro), cars, and pedestrian trips. The prediction of mode of transportation for flows was made while using a variation of the model published in Ref. [42], aiming at solving a limitation of the original method, namely, working with morning commuting trips for individual devices. We have updated the model to estimate modal partitions in flows between areas of the city at several periods of the day. The updated model consists of solving the following optimization problem:

$$\min_{A,B} \|W - (L \odot A) \times B\|_F^2,$$

where:

- W is a waypoint matrix, where every column represents a directed flow between two areas of the city, and every rows represents a tower. Thus, each cell w_{ij} contains the number of times that tower i appears in the trajectories in flow j.
- A and B are positive low-rank matrices that express the associations between k latent dimensions and each tower (A) and flow (B).
- L is a k-rank labeling matrix where l_{ij} is 1 if tower i is associated to mode of transportation j, 0 otherwise. This labeling enables aligning latent dimensions with mode of transportation usage in a semi-supervised way, as some towers are strictly associated to specific modes of transportation due to urban infrastructure surrounding it (see Ref. [42] for details).
- $\odot$ is the Hadamard product operator.

The optimization problem can be solved while using multiplicative updates. After solving the problem, the matrices A and B contain the association between towers and flows with the k latent dimensions. These associations are interpreted as modal partitions per tower and per flow.

In this work, we average the flows per period of the day (morning peak 1, morning peak 2, afternoon valley, afternoon peak, night valley, night) in a single representative day, with a focus on working days.

4.2. General Layout

Modalflow works as a set of coordinated visualizations displayed in a notebook-style layout [43], i.e., visualizations are stacked in display and usage order from top to bottom, without constraining the layout to fit into a screen. Gleichner categorizes comparison layout strategies as juxtaposition, superposition, and explicit encoding [44]. We chose to use juxtaposition, because using overlay would have resulted in a cluttered view, and explicit encoding would have removed the data

context [45]. Each layout element is a view from the system where the different datasets are visualized in juxtaposition to ease comparison across views. Each visualization shows different dimensions of the data and acts in coordination with the rest: when a selection is created in one view of the system, it extends to the other dataset and views of the system. Note that our system is notebook-style and not a notebook, as the user interaction only happens through the coordinated visualizations.

4.3. Color Coding

To encode modal partition throughout views (R3), we decided to use a color mapping, as color is the only visual channel that can be applied invariably across representations. Using the constraint that in urbanism the complexity of transportation is categorized into three main modes (public, private and non-motorized) the modal partition could be mapped to a ternary color scale, as shown in Figure 2. Non-motorized is mapped to green, public to cyan, and private to magenta, mixed flows being a proportional combination of these hues. Thus, in the case of a totally balanced mode split, the resulting color will be gray.

To approximate the desired quality of perceptual uniformity, where all of the points have the same luminance value, we used the HSLuv color space [46], which is an HSL model version of the perceptually uniform CIELUV color space. However, as the color coding is naturally defined in a three-channel additive color model as RGB (each channel mapped to a transport mode), we devised the color transformation pipeline RGB $\rightarrow$ HSL $\rightarrow$ HSLuv $\rightarrow$ RGB to transit from a non-perceptually uniform to a (at least more) perceptually uniform RGB color code.

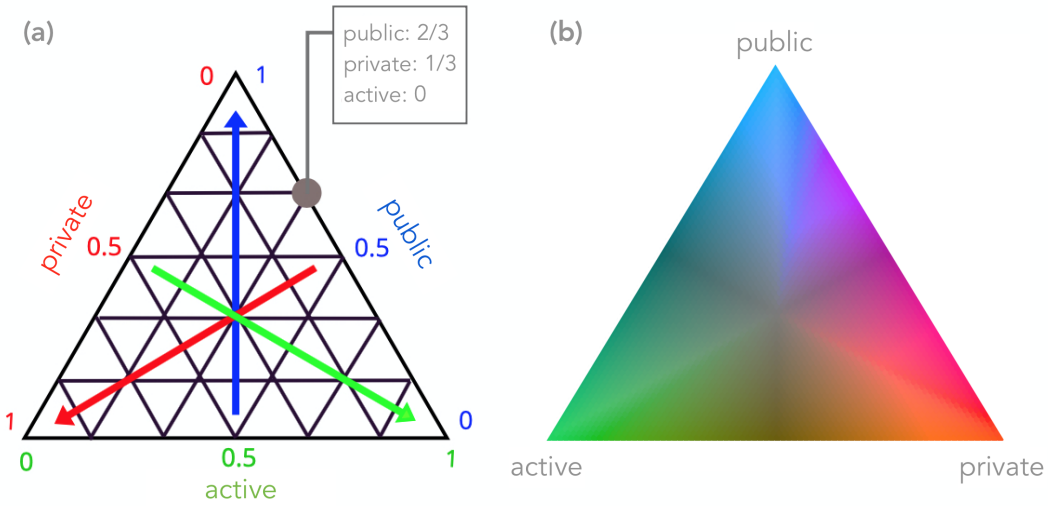

Figure 2. Ternary color scale construction schema (**a**) and resulting colormap (**b**) used to encode mode split across views.

4.4. Ternary Plot

Ternary plots show the distribution of points in three dimensions that sum to a constant, as is the case of modal partition for flows. We encode the flows as circles and optionally encode magnitude in its area. Positioning OD flows on the ternary plot can show particular biases of the surveying or inference mechanism toward certain modes, especially when used in combination with range selections in other dimensions. Additionally, when outgoing flows are aggregated by municipality, it becomes direct to compare mean modal behaviour and observe its change over the day.

4.5. Scatterplot

For a view over the abstract (non-spatial) distribution of data, we customized a scatterplot schema over flow magnitude and distance dimensions (R2). Every OD flow is represented as a clear circle of

fixed or flow magnitude-weighted radius in the two-dimensional space. Selections can be made by dragging over the plot, thus selecting a distance-flow magnitude pair range, i.e., a rectangular area. Many different selections can be maintained at the same time, which can overlap or be embedded into each other. Statistics are calculated over a new selection (number of trips, flows, total distance, and portion of the whole) and annotated on the side.

Selection ranges are also "draggable" while using right and left arrow keys. This interaction moves the selection range over the distance dimension, allowing to see the user to focus on another visualization and observe how distance correlates with the flow distribution on another view of the system.

While testing the scatterplot, we discovered that normal range selection proved to be not effective for comparing different datasets, as range selections in one dataset would probably lead to an empty selection in the other one due to the very different distributions of data. Thus, to comply with R1, we developed a distribution-aware selection tool.

As the distribution of the examined datasets proved to be different over this space, selections that were made over one dataset usually ended in an empty selection over the other one, making a the comparative approach ineffective. For this reason, we developed a distribution-aware selection tool that translates a selection over one distribution into a more alike selection over another distribution, as shown in Figure 3. It is a linear transformation of the XY space based on data distribution. This technique can be generalized to different datasets; however, a decision has to be made over the formal nature of their distribution to define the linear mapping. In this case, preliminary visualization experiments showed a power law-type distribution; thus, a measure of its powerlaw had to be taken into account (for a normal distribution, for example, its median and variance should determine the outcome mapping).

In order to determine the mapping for the distribution-aware tool, the method is as follows: (1) for each dimension, sort the data according to that dimension and determine the threshold at which the highest ranking elements are equal in value to all the others; and, (2) this threshold is then the value of the scale factor k_i for dataset i. The mapping of coordinate x_j from coordinate x_i is given by

$$x_j = (x_i/k_i)k_j.$$

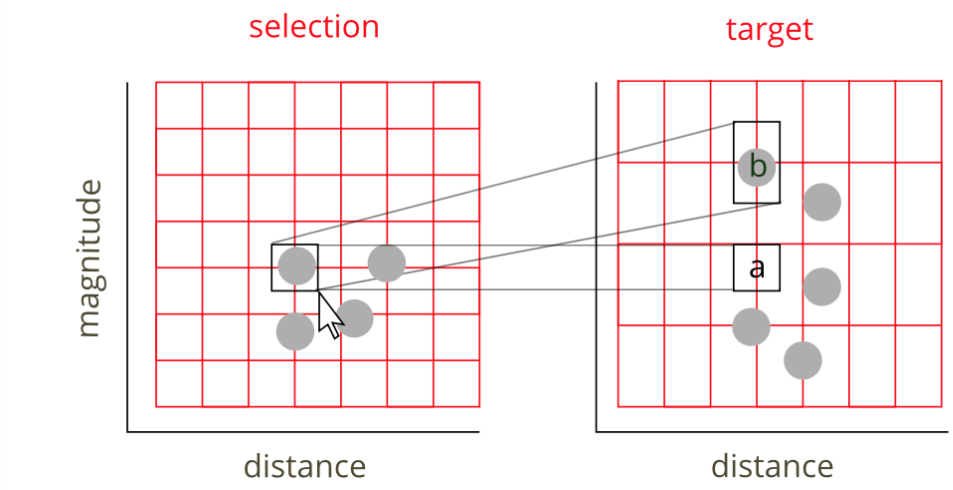

Figure 3. Selection over different datasets in the scatterplot is facilitated by our distribution-aware selection tool. A range selected in one scatterplot is not merely transposed to *a*, which would probably end in an empty selection, but linearly transformed into *b*, according to the grid automatically defined based on the distribution of each dataset.

For the particular case of mobility flows, we deemed that using this technique only in the magnitude dimension was the most sensible choice, because using distribution-aware selection over distance would change the original meaning of a selection.

4.6. OD Matrix

We present two optimizations, in terms of our design requirements, for the OD matrix: (1) a disaggregated layout option for showing zonal flows (R1); and, (2) a color encoding for mode split following the color map previously defined (R3).

In the disaggregated layout, each cell of the municipal OD matrix is horizontally divided by the number of zones in the destination municipality and vertically by the number of zones in the origin municipality, thus producing equally sized submatrices. At the cost of not having equally sized cells in the submatrices, this procedure preserves the space filling layout and the visual correspondence between aggregation states.

4.7. Flow Map with Sinusoidal Encoding

The flow map is the overlaying of a graph (where nodes are georeferenced points and flows are directed edges) on top of a geographic map. Its function is to reveal geospatial patterns, like sinks and sources of flows. However, as representing the flows for a whole city leads to a lot of clutter and low readability in the direction of flows, we developed a novel encoding for flows that uses a sinusoidal curve rather than an arrow or arc.

Jane Jacobs conceptualized transport as "a set of networks stitching together spaces and places, from which locations naturally emerge" [47]. This metaphor of transport weaving the fabric of the city is evocative of a graph-like visualization of flows. Through it, we were inspired to design a new encoding to facilitate distinguishing incoming from outgoing flows, particularly over a geographic layout. Figure 4 schematically shows the two most common encoding alternatives for links in a graph, followed by our novel sinusoidal encoding. This encoding was designed, so that it brings a new symmetry into perspective, where directed links that go through a node get immediately divided into incoming and outgoing in the vertical dimension, which makes for an intuitive way to select them, as shown in Figure 4.

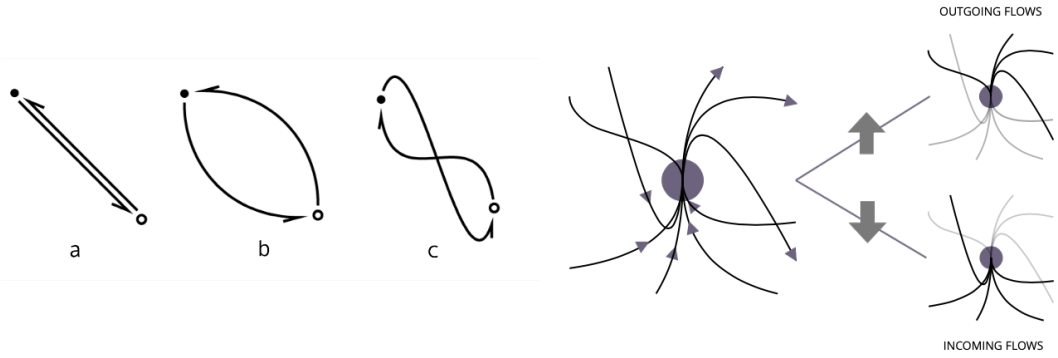

Figure 4. Sinusoidal encoding design space and working schema. Left: Bidirectional flow encoding alternatives: (**a**) straight; (**b**) arched; and, (**c**) sinusoidal, which projects a third dimension into perspective, differentiating in- and outflows in the vertical dimension. Right: schematic view of incoming and outgoing flows at a certain node. Flow direction becomes quickly distinguishable thanks to the sinusoidal flow encoding.

The sinusoidal encoding is achieved using a centripetal Catmull–Rom spline for each directed node-pair. The spline is defined by control points P_0, P_1, P_2, and P_3, where P_1 and P_2 are the position of the start node and end node, correspondingly, and

$$P_0 = (P_1x, P_1y + k * distance(P_1, P_2)),$$

$$P_3 = (P_2x, P_2y - k * distance(P_2, P_1)),$$

where k is a constant that ponders the vertical elongation of the curves.

As this new edge drawing technique does not get rid of the clutter, but rather tries to improve its visual organization, two known techniques to improve readability were also implemented in the rendering of flows: edge sorting (edges representing bigger flows are drawn on the front) and transparency (edges that represent less flow magnitude are less opaque).

4.8. Edge Bundling

An edge bundled view of the flows is also available [48]. This encoding for the flow map is used in order to produce visual structures with coarser features to show major trends in the data, particularly for mode-filtered dataset comparison (R1), and so, this is the only encoding where the modal color scale is not used, but a radial color scale here is used to encode bundled flow direction.

Taking into account our data-set specificity, we selected the most appropriate edge bundling technique, which allies computation simplicity and a sufficient visual aggregation results, which balances data distortion and visual accuracy [16]. For this reason, we applied the KDEEB technique [49].

5. Results

Here, we describe the insights that we, as researchers and users of the system, learned by comparing pervasive data to survey data, in a general sense, and about the particular inference model behind the XDR dataset.

The scatterplot reveals that pervasive data capture shorter and longer flows better than the survey, which has a slightly narrower distance range and is also biased towards mid-distance flows, as shown on Figure 5. This difference can have a rather important effect on the interpretation of data: a steeper, long-tailed (powerlaw) distribution of flows suggests a scale-invariant network phenomenon, while the flatter survey distribution does not, and thus surveys do not capture the full complexity of mobility.

XDR has almost total municipal origin-destination pair sampling coverage for a given period, whereas survey data do not, as shown through the OD Matrix in Figure 6. Additionally, there is an observable pattern of sampling in survey data: short active flows are unintentionally oversampled (showing in the bright diagonal). However, when flows are expanded, these flows are downplayed and overshadowed by expanded public and private mode longer-distance flows.

The geographic view of the flows is quite dissimilar for both datasets, exhibiting rather different features, as can be observed in Figure 7. Pervasive data show notorious change in behaviour during the day: in the morning, peripheral flows are directed towards the center and northeastern part of the city (the small peaks appearing on these nodes); flows in the afternoon show roughly the opposite pattern, but also an emergence of active flows around certain parts of the city; at night, short flows become totally predominant across the city. On the other hand, survey data, apart from density changes, do not exhibit a coherent structure in each period or in time, as its most prominent features are chaotic mid- and long-distance flows.

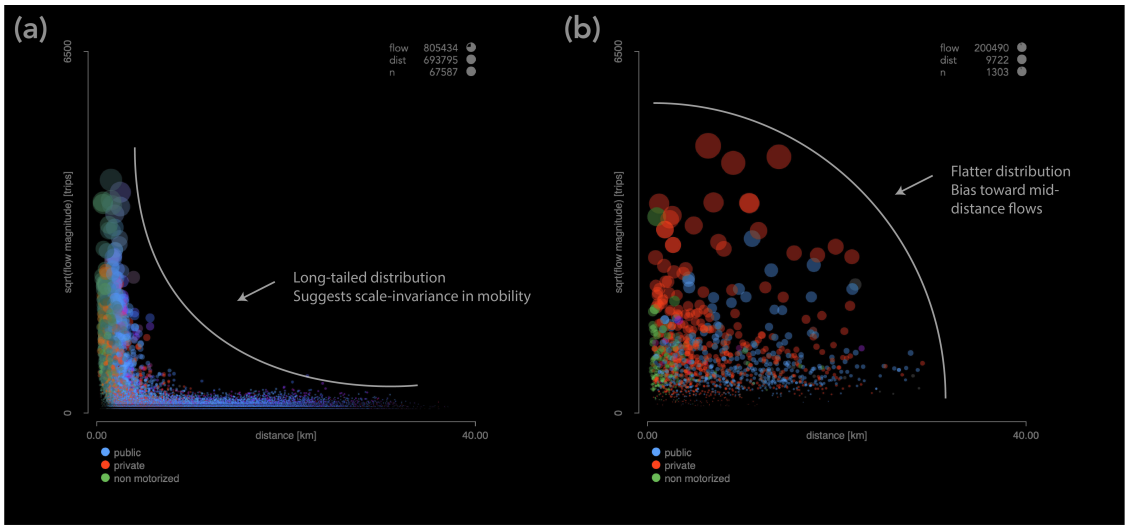

Figure 5. Pervasive (**a**) and survey (**b**) flows at night time. Each circle is an origin-destination pair, its area the flow magnitude, with color coded mode split. Highlighted are the different overall shapes present in their distribution.

The high density and large flows inside neighbors suggests that network-based analysis (e.g., community analysis) could be possibleusing XDR data, while, with the travel survey, a network structure does not appear so clearly, as shown by the OD matrix in Figure 8. Also, the fact that pervasive data resembles more a powerlaw mean that it could be a much better fitted by a model, e.g., the gravitational model [50], and so it would be a better input choice for a trained model, and for network-based analysis.

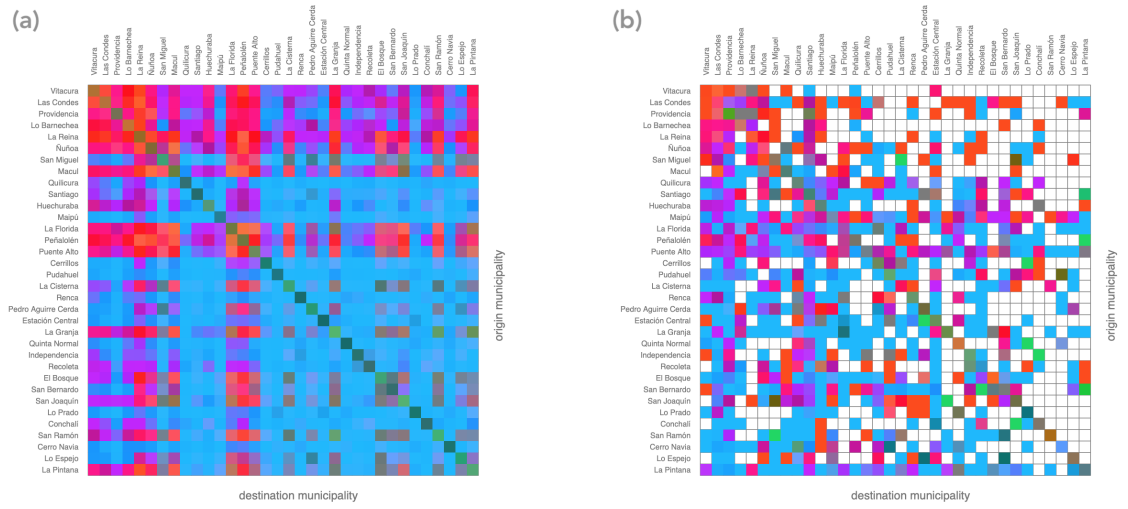

Figure 6. XDR (**a**) and survey (**b**) flows at morning peak 1 at municipal aggregation level. Even at a time period where they show a similar total flow magnitude, the sampling density is different.

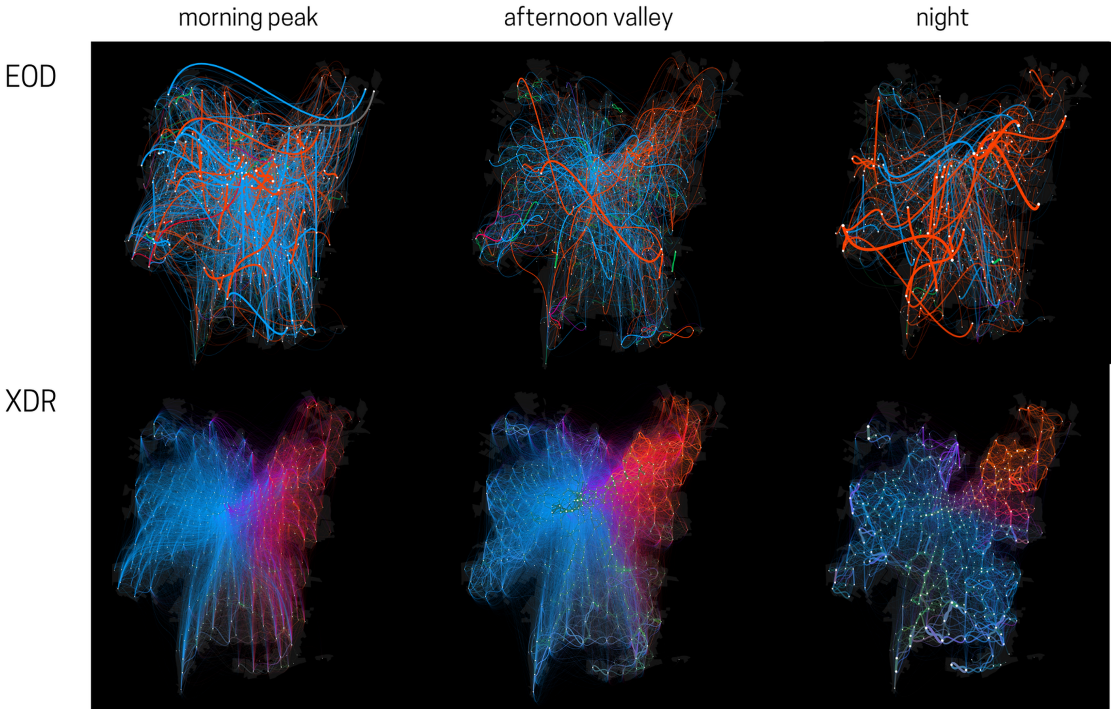

Figure 7. One day of mobility according to a travel survey (EOD) and a pervasive (XDR) dataset, visualized as flow maps on Modalflow.

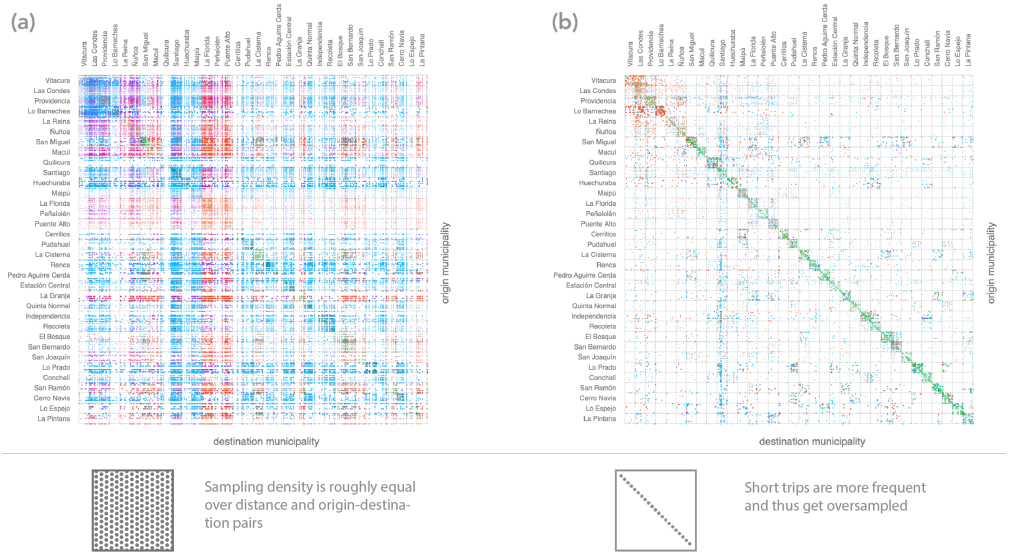

Figure 8. Zone-level OD matrix of XDR (**a**) and survey (**b**) flows for the night valley period. Sparser measures and an over representation of intramunicipal flows can be appreciated in the survey.

Regarding modal distribution, the results are less general and more specific to the particular mode of transportation inference model. This is due to the fact that the visualization reveals many features that point to an over-fitted model, which is explained by the extended context in which the model was applied (and not prepared for).

A quick observation of bundled flows on Figure 9 a particularly harsh contrast between private mode distributions, with flows in the northeastern corner of the city heavily charged towards private car usage and almost none in the rest of the city. Public transport is known to be less predominant in the northeastern area of the city, which leads to the hypothesis that the current inference model may be drifting toward extreme border conditions that are based on a small spatial trend. By looking through different periods, as in Figure 7 we observe that this spatial partitioning of the city is stable, but it has

a dynamic component, as sometimes the private mode cluster changes its geographical reach (private is displaced to the southeast at Night Valley, but then goes back to its most common post).

The system is also able to reveal a strange pattern in the active flows: although they look evenly sparsed geographically, a suspicious trend appears in the relation of the distance-modal split distribution of active trips. This analytical task is achieved by dragging a narrow selection range horizontally in the scatterplot, while looking at the ternary plot. By using this interaction technique, the modal progression of the partial flow distribution as distance grows can be observed, which reveals that inferred active trips dominate the very short distance up to the limit of around 2 km, where they abruptly recede, but at all distances keeping at least one flow that is purely attributed to active mobility (a certain artifact of the model). By doing the same experiment, a heavier correlation of inferred private trips with distance is observed (compared to the survey data).

Finally, by looking at the compared modal distribution of flows that are aggregated by origin municipality in the ternary plot (see Figure 10), an overall trend towards a bigger proportion of active and public mode is observed in the XDR data. This result can be attributed to a bias in the model, but also to the fact that pervasive data are more capable of detecting these trips than the traditional survey.

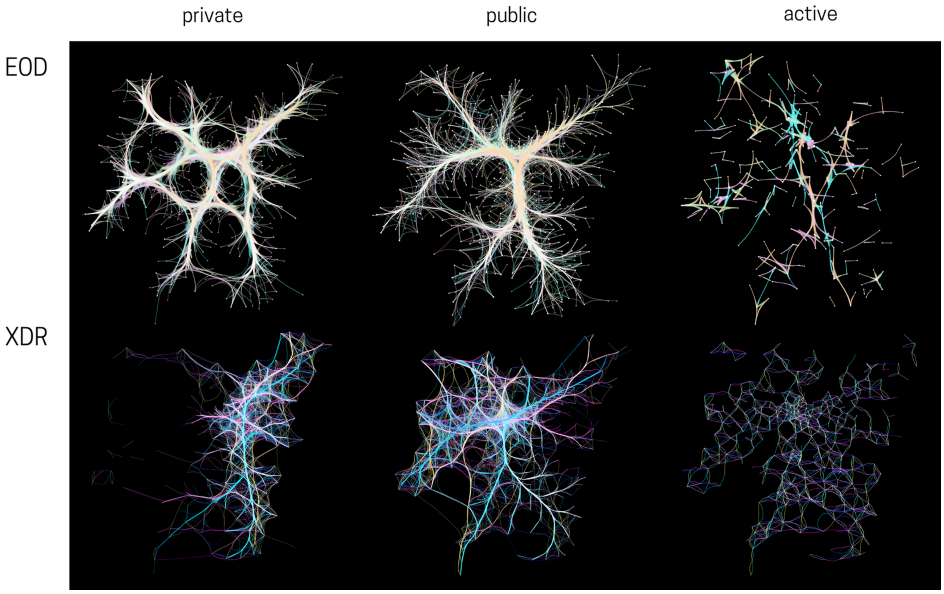

Figure 9. Comparison of survey (**EOD**) and pervasive (**XDR**) data with geographic bundling by mode of transportation.

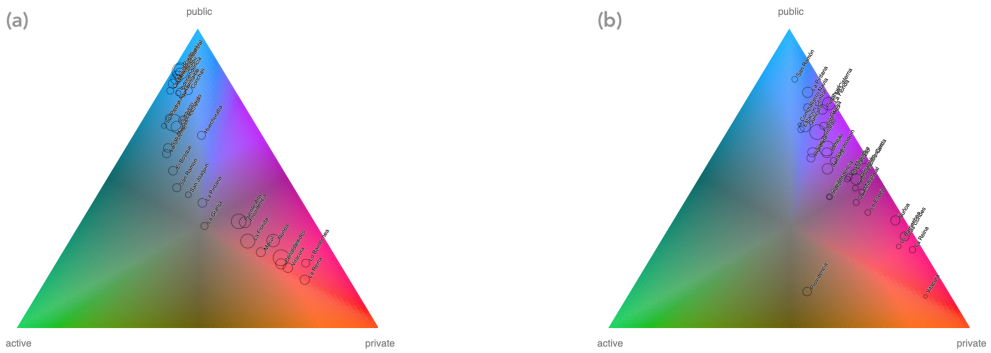

Figure 10. XDR (**a**) and survey (**b**) flows aggregated by origin municipality positioned on the ternary plot according to mean trip mode split (also encoded by background color), with radius encoding total magnitude.

6. Discussion

We showed Modalflow to domain experts and potential users working in Santiago through an asynchronous qualitative evaluation process lasting a month. Each week, a new view of the system would be uploaded to the demo page and iterated according to modifications that were suggested by the freely commenting participants in the online discussion room. In some occasions, the participants would even mock-up a new solution and present it to the group of their own will. In total, eight people (four visualization/urban studies students and four urban planners from the private sector) voluntarily participated throughout the whole process. After this, a one-session workshop was arranged with governmental transport authorities, where five experts were shown the final system. Here, we report on the results.

Early feedback was incorporated into the design, like the use of black background, only three modes of transportation and the size of the visual elements, which can be switched between fixed and variable (dependant on encoded flow magnitude). The sinusoidal encoding of flows was deemed to be beautiful, but a bit confusing and using the more familiar arc encoding was suggested. We explain this comment based on familiarity but we are aware that a formal evaluation is needed. Interestingly, the sinusoidal encoding was more readable when dealing with XDR, rather than survey data, arguably due to the artificial modal clustering computed for the former.

Admittedly, the domain experts were deterred at a later stage by the modal distribution of the mobile phone network data, which, being still undergoing research, did not meet their expectations. It seemed hard for them to abstract the visualization tool from the data being visualized, expecting to see results rather than a tool to evaluate results by themselves. Conversely, the group of experts from the private sector received the tool with much more interest and excitement. Even though both domain expert profiles work on similar problems, experts working in the public sector seemed to be more focused on strict results that could be readily applied, whereas experts in the private sector were more open to new methodologies and how these visualizations "create" a new reality for the data that can be used to strongly communicate and innovate. This difference has implications in how we define and apply methodologies to enable collaborative work between transportation and data science, using visualization as intermediary between disciplines [39].

The distribution-aware selector tool in the scatterplot was received as an interesting feature and its implications could be further developed, attending to the idea of an "equivalence principle" between flows of different datasets. Such an endeavour could be an interesting area of research for modelling and mixed data analysis. This also poses new questions regarding the layout of the system. Modalflow was designed with linked, juxtaposed visualizations, where the selector tool provided a way to link data semantics. However, there exist other ways of composing visualizations in layouts, such as superimposition, overloading and nesting [51], which would imply a different implementation of the selection, while maintaining the data-semantic approach. These layouts will be studied in future work. Because of the notebook layout, they are not exclusive—for instance, the OD matrices may be juxtaposed, whereas the flow maps may be partitioned or nested, enabling an advanced comparison and interaction with cross-origin datasets.

Other shortcomings of the design that were noted by experts were that the color encoding used is not color-blind safe (making a three-component color scale to be inclusive is a challenge by itself, which we propose as a future line of research) and that defining a three-fold categorical variable as basis for the system seems like a limited choice, as some urban planners may want to disaggregate the modes of transportation into four, or the usage of another categorical variable may need more categories (for instance, income quintiles). From a technical standpoint, our choice presents a trade-off between an overview and a detailed view of the datasets. The use of a notebook-style layout allows to dive into a deeper level of details through adding more visualizations. These improvements will be addressed in future versions of the system. Still, the use of three categories seems to be enough to already tackle on-going problems in mobility, for instance, there are several gaps in mobility that can be explored with two or three variables [52]; thus, Modalflow would allow for urban planners to

uncover not only biases in methodological workflows, but also gaps in behaviour that are commonly hidden in traditional analyses and visualizations [31,32].

7. Conclusions

In this paper, we have presented the Modalflow system for visualizing modal flow data in mobility, with an emphasis comparing cross-origin flow data sources. The research contribution of our paper consists of two main components: a set of considerations regarding the nature of pervasive and traditional data sources, which we then used to determine the requirements for our system design, and a set novel visualization techniques to fulfill those requirements, including a distribution-aware selection tooltip and a bidirectional flow encoding. We validated our approach with a case study while using real-world cross-origin datasets: XDR-inferred and traditional survey data of the city of Santiago.

Our aim was to create a tool that could help domain experts in urbanism to assess new datasets from non-traditional sources and hopefully incorporate them into their workflow, at the same time fostering interaction between data science and urbanism, as we saw this as an important gap to be filled by the visualization community. Having presented the tool to different groups of experts, the experts emphasised the importance of the problem addressed in this research and recognized Modalflow as a notable tool to communicate the model results across disciplines.

While there is room for improving and extending each of the techniques presented in this paper (geographical flow map with sinusoidal encoding/edge bundling, distance/magnitude scatterplot, disaggregated origin-destination matrix, and ternary modal plot), we other important aspects of future work: (1) exploring the temporal dimension: some visualizations developed point to new ways of looking at multivariate flows that could be further researched (e.g., in the ternary plot, when flows are aggregated by origin municipality, it gives a condensed image of the modal partition of the city; enhancing it with the curves that were described by each municipality in the modal space through time, an interesting research about daily mobility behaviour of cities could be conducted). (2) Finding ways to visualize and compare accessibility (i.e., the analytical dimension of flows in urbanism that considers the time that it takes to make a certain trip and other physical constraints) between datasets, a dimension that was not considered in our current work, which adds a new layer of complexity to the data. (3) Using the system to extend our study to different sources of pervasive data, thus developing a better understanding of their differential biases, and to leverage human critical thinking in the task of finding ways of making cross-origin datasets complement each other. We think that this paper stresses the importance of visualization as a critical tool from a human-in-the-loop perspective, particularly in complex scenarios, where no benchmark or ground truth is available, such as urban mobility.

Author Contributions: Conceptualization, I.P.-M., E.G.-G.; methodology, I.P.-M., E.G.-G., M.J.L., C.H.; software, I.P.-M.; validation, I.P.-M.; formal analysis, I.P.-M.; investigation, I.P.-M., E.G.-G.; resources, E.G.-G.; data curation, E.G.-G.; writing–original draft preparation, I.P.-M.; writing–review and editing, I.P.-M., E.G.-G., M.J.L., C.H.; visualization, I.P.-M.; supervision, E.G.-G.; project administration, I.P.-M., E.G.-G.; funding acquisition, E.G.-G. All authors have read and agreed to the published version of the manuscript.

Funding: I.P.-M. and E.G.-G. were partially funded by CONICYT Fondecyt de Iniciación project #11180913.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Andrienko, G.; Andrienko, N.; Chen, W.; Maciejewski, R.; Zhao, Y. Visual analytics of mobility and transportation: State of the art and further research directions. *IEEE Trans. Intell. Transp. Syst.* **2017**, *18*, 2232–2249. [CrossRef]
2. Zegras, P.C.; Li, M.; Kilic, T.; Lozano-Gracia, N.; Ghorpade, A.; Tiberti, M.; Aguilera, A.I.; Zhao, F. Assessing the representativeness of a smartphone-based household travel survey in Dar es Salaam, Tanzania. *Transportation* **2018**, *45*, 335–363. [CrossRef]
3. Wolf, J.; Loechl, M.; Thompson, M.; Arce, C. Trip rate analysis in GPS-enhanced personal travel surveys. In *Transport Survey Quality and Innovation*; Emerald Group Publishing Limited: Bingley, UK, 2003; pp. 483–498.

4. Blondel, V.D.; Decuyper, A.; Krings, G. A survey of results on mobile phone datasets analysis. *EPJ Data Sci.* **2015**, *4*, 10. [CrossRef]
5. Nitti, M.; Pinna, F.; Pintor, L.; Pilloni, V.; Barabino, B. iABACUS: A Wi-Fi-Based Automatic Bus Passenger Counting System. *Energies* **2020**, *13*, 1446. [CrossRef]
6. Ruiz Sánchez, T.; Mars Aicart, M.D.L.; Arroyo-López, M.R.; Serna, A. Social networks, big data and transport planning. *Transp. Res. Procedia* **2016**, *18*, 446–452. [CrossRef]
7. Chen, W.; Guo, F.; Wang, F.Y. A survey of traffic data visualization. *IEEE Trans. Intell. Transp. Syst.* **2015**, *16*, 2970–2984. [CrossRef]
8. Yang, Y.; Dwyer, T.; Goodwin, S.; Marriott, K. Many-to-many geographically-embedded flow visualisation: An evaluation. *IEEE Trans. Vis. Comput. Graph.* **2017**, *23*, 411–420. [CrossRef]
9. Wood, J.; Dykes, J.; Slingsby, A. Visualisation of origins, destinations and flows with OD maps. *Cartogr. J.* **2010**, *47*, 117–129. [CrossRef]
10. McNeill, G.; Hale, S.A. Generating tile maps. In *Computer Graphics Forum*; Wiley Online Library: Hoboken, NJ, USA, 2017; Volume 36, pp. 435–445.
11. Von Landesberger, T.; Brodkorb, F.; Roskosch, P.; Andrienko, N.; Andrienko, G.; Kerren, A. Mobilitygraphs: Visual analysis of mass mobility dynamics via spatio-temporal graphs and clustering. *IEEE Trans. Vis. Comput. Graph.* **2015**, *22*, 11–20. [CrossRef]
12. Zhu, X.; Guo, D. Mapping large spatial flow data with hierarchical clustering. *Trans. GIS* **2014**, *18*, 421–435. [CrossRef]
13. Guo, D.; Zhu, X. Origin-destination flow data smoothing and mapping. *IEEE Trans. Vis. Comput. Graph.* **2014**, *20*, 2043–2052. [CrossRef] [PubMed]
14. Graser, A.; Schmidt, J.; Roth, F.; Brändle, N. Untangling origin-destination flows in geographic information systems. *Inf. Vis.* **2019**, *18*, 153–172. [CrossRef]
15. Holten, D.; Van Wijk, J.J. Force-directed edge bundling for graph visualization. In *Computer Graphics Forum*; Wiley Online Library: Hoboken, NJ, USA, 2009; Volume 28, pp. 983–990.
16. Lhuillier, A.; Hurter, C.; Telea, A. FFTEB: Edge bundling of huge graphs by the Fast Fourier Transform. In Proceedings of the 2017 IEEE Pacific Visualization Symposium (PacificVis), Seoul, Korea, 18–21 April 2017; pp. 190–199, [CrossRef]
17. Zhou, H.; Xu, P.; Yuan, X.; Qu, H. Edge bundling in information visualization. *Tsinghua Sci. Technol.* **2013**, *18*, 145–156, [CrossRef]
18. Holten, D. Hierarchical Edge Bundles: Visualization of Adjacency Relations in Hierarchical Data. *IEEE Trans. Vis. Comput. Graph.* **2006**, *12*, 741–748, [CrossRef] [PubMed]
19. Hurter, C.; Puechmorel, S.; Nicol, F.; Telea, A. Functional Decomposition for Bundled Simplification of Trail Sets. *IEEE Trans. Vis. Comput. Graph.* **2018**, *24*, 500–510, [CrossRef]
20. Wang, Y.; Xue, M.; Wang, Y.; Yan, X.; Chen, B.; Fu, C.; Hurter, C. Interactive Structure-aware Blending of Diverse Edge Bundling Visualizations. *IEEE Trans. Vis. Comput. Graph.* **2020**, *26*, 687–696, [CrossRef]
21. Borgo, R.; Kehrer, J.; Chung, D.H.; Maguire, E.; Laramee, R.S.; Hauser, H.; Ward, M.; Chen, M. *Glyph-Based Visualization: Foundations, Design Guidelines, Techniques and Applications*; Eurographics (STARs): Aire-la-Ville, Switzerland, 2013; pp. 39–63.
22. Ma, Y.; Lin, T.; Cao, Z.; Li, C.; Wang, F.; Chen, W. Mobility viewer: An Eulerian approach for studying urban crowd flow. *IEEE Trans. Intell. Transp. Syst.* **2016**, *17*, 2627–2636. [CrossRef]
23. Andrienko, G.; Andrienko, N.; Fuchs, G.; Wood, J. Revealing patterns and trends of mass mobility through spatial and temporal abstraction of origin-destination movement data. *IEEE Trans. Vis. Comput. Graph.* **2017**, *23*, 2120–2136. [CrossRef]
24. Pérez-Messina, I.; Graells-Garrido, E. *Visualizing Transportation Flows with Mode Split Using Glyphs*; EuroVis 2019—Short Papers; Johansson, J., Sadlo, F., Marai, G.E., Eds.; The Eurographics Association: Genoa, Italy, 2019. [CrossRef]
25. Zeng, W.; Fu, C.W.; Arisona, S.M.; Qu, H. Visualizing interchange patterns in massive movement data. In *Computer Graphics Forum*; Wiley Online Library: Hoboken, NJ, USA, 2013; Volume 32, pp. 271–280.
26. Zeng, W.; Fu, C.W.; Müller Arisona, S.; Erath, A.; Qu, H. Visualizing Waypoints-Constrained Origin-Destination Patterns for Massive Transportation Data. In *Computer Graphics Forum*; Wiley Online Library: Hoboken, NJ, USA, 2016; Volume 35, pp. 95–107.

27. Wesolowski, A.; Stresman, G.; Eagle, N.; Stevenson, J.; Owaga, C.; Marube, E.; Bousema, T.; Drakeley, C.; Cox, J.; Buckee, C.O. Quantifying travel behavior for infectious disease research: A comparison of data from surveys and mobile phones. *Sci. Rep.* **2014**, *4*, 5678. [CrossRef]
28. Stopher, P.; Shen, L. In-depth comparison of global positioning system and diary records. *Transp. Res. Rec.* **2011**, *2246*, 32–37. [CrossRef]
29. Safi, H.; Assemi, B.; Mesbah, M.; Ferreira, L. An empirical comparison of four technology-mediated travel survey methods. *J. Traffic Transp. Eng.* **2017**, *4*, 80–87. [CrossRef]
30. Son, S.; Khattak, A.; Wang, X.; Agnello, P.; Chen, J.Y. Quantifying Key Errors in Household Travel Surveys: Comparison of Random-Digit-Dial Survey and Address-Based Survey. *Transp. Res. Rec.* **2013**, *2354*, 9–18. [CrossRef]
31. Kwan, M.P. Feminist visualization: Re-envisioning GIS as a method in feminist geographic research. *Ann. Assoc. Am. Geogr.* **2002**, *92*, 645–661. [CrossRef]
32. D'Ignazio, C.; Klein, L.F. Feminist data visualization. In Proceedings of the Workshop on Visualization for the Digital Humanities (VIS4DH), Baltimore, MD, USA, 24 October 2016.
33. Dörk, M.; Feng, P.; Collins, C.; Carpendale, S. Critical InfoVis: Exploring the politics of visualization. In Proceedings of the CHI'13 Extended Abstracts on Human Factors in Computing Systems, Paris, France, 27 April–2 May 2013; pp. 2189–2198.
34. Hall, P.; Heath, C.; Coles-Kemp, L. Critical visualization: A case for rethinking how we visualize risk and security. *J. Cybersecur.* **2015**, *1*, 93–108. [CrossRef]
35. Drucker, J. Humanities approaches to graphical display. *Digit. Humanit. Q.* **2011**, *5*, 1–21.
36. Alexander, L.; Jiang, S.; Murga, M.; González, M.C. Origin–destination trips by purpose and time of day inferred from mobile phone data. *Transp. Res. Part C Emerg. Technol.* **2015**, *58*, 240–250. [CrossRef]
37. Graells-Garrido, E.; Ferres, L.; Caro, D.; Bravo, L. The effect of Pokémon Go on the pulse of the city: A natural experiment. *EPJ Data Sci.* **2017**, *6*, 23. [CrossRef]
38. Subdirección General de Estadísticas Sociodemográficas. Estudio EM-1 de Movilidad a Partir de la Telefonía Móvil. Instituto Nacional de Estadísticas. 2020. Available online: https://www.ine.es/experimental/movilidad/exp_em1_proyecto.pdf (accessed on 13 November 2020).
39. Graells-Garrido, E.; Peña-Araya, V.; Bravo, L. Adoption-Driven Data Science for Transportation Planning: Methodology, Case Study, and Lessons Learned. *Sustainability* **2020**, *12*, 6001. [CrossRef]
40. Pérez-Messina, I. *Modalflow Demo*. 2020. Available online: http://www.baltazarperez.com/modalflow-demo/ (accessed on 13 November 2020).
41. Universidad Alberto Hurtado, Observatorio Social. Encuesta Origen Destino Santiago 2012 (Informe Ejecutivo). Available online: http://www.sectra.gob.cl/biblioteca/detalle1.asp?mfn=3253 (accessed on 30 April 2020).
42. Graells-Garrido, E.; Caro, D.; Parra, D. Inferring modes of transportation using mobile phone data. *EPJ Data Sci.* **2018**, *7*, 49. [CrossRef]
43. Kery, M.B.; Radensky, M.; Arya, M.; John, B.E.; Myers, B.A. The story in the notebook: Exploratory data science using a literate programming tool. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, Montreal, QC, Canada, 21–26 April 2018; pp. 1–11.
44. Gleicher, M.; Albers, D.; Walker, R.; Jusufi, I.; Hansen, C.D.; Roberts, J.C. Visual comparison for information visualization. *Inf. Vis.* **2011**, *10*, 289–309, [CrossRef]
45. Gleicher, M. Considerations for visualizing comparison. *IEEE Trans. Vis. Comput. Graph.* **2017**, *24*, 413–423. [CrossRef] [PubMed]
46. Boronine, A. *HSLuv*. 2020. Available online: https://www.hsluv.org (accessed on 13 November 2020).
47. Maheshwari, T.; Fourie, P.J.; van Eggermond, M.A. Transportation flows in future cities. In *Future Cities Laboratory: Indicia 02*; Lars Müller Publishers: Baden, Switzerland, 2019; pp. 119–217.
48. Scheepens, R.; Hurter, C.; van de Wetering, H.; van Wijk, J. Visualization, Selection, and Analysis of Traffic Flows. *IEEE Trans. Vis. Comput. Graph.* **2015**, *22*. [CrossRef] [PubMed]
49. Hurter, C.; Ersoy, O.; Telea, A. Graph Bundling by Kernel Density Estimation. *Comput. Graph. Forum* **2012**, *31*, 435–443, [CrossRef]

50. Beiró, M.G.; Bravo, L.; Caro, D.; Cattuto, C.; Ferres, L.; Graells-Garrido, E. Shopping mall attraction and social mixing at a city scale. *EPJ Data Sci.* **2018**, *7*, 28. [CrossRef]
51. Javed, W.; Elmqvist, N. Exploring the design space of composite visualization. In Proceedings of the 2012 IEEE Pacific Visualization Symposium, Songdo, Korea, 28 February–2 March 2012; pp. 1–8.
52. Graells-Garrido, E.; Meta, I.; Serra-Buriel, F.; Reyes, P.; Cucchietti, F.M. Measuring Spatial Subdivisions in Urban Mobility with Mobile Phone Data. In Proceedings of the Companion Proceedings of the Web Conference 2020, Taipei, Taiwan, 20–24 April 2020; pp. 485–494.