CENTENNIAL PAPERS

Standing Committee on Urban Transportation Data and Information Systems (ABJ30)
Stacey Bricka, Chair

The Future of Urban Data

MARCELO SIMAS, PHD, *Westat*
LETA HUNTSINGER, PHD, PE, *WSP*
STACEY BRICKA, PHD, *MacroSys Research and Technology*

INTRODUCTION

Formally recognized as a committee by TRB in the 1990s, the TRB Committee on Urban Transportation Data and Information Systems (ABJ30) is interested in the design, collection, analysis, and reporting of transportation supply and demand data and the information systems needed to support the application of that data in urban and metropolitan transportation planning efforts. In particular, the committee is interested in:

- New and innovative techniques for measuring and monitoring the performance of metropolitan transportation systems;
- Impacts associated with changes in demographic and urban travel behavior characteristics;
- Effective use of primary (household and other transportation surveys) and secondary (census and other federal, state, local and passive data sources) data;
- Advancements in information systems and information technology for improved dissemination and sharing of knowledge about metropolitan transportation systems and urban travel behavior, including the role of big data; and
- Common standards and appropriate recommendations to support the interchange and archiving of information and data.

As noted in the 2018-2020 TRB Urban Data and Information Systems (ABJ30) triennial strategic plan, we are entering a critical juncture in the area of urban transportation data and information systems.

- On the supply side, system usage as expressed through traffic volume, speed and congestion data are obtained using a mix of old (e.g., loop detectors, sensors, video cameras) and new (e.g., Bluetooth, GPS, connected and autonomous vehicle (CAV), transportation network company (TNC)) technologies. This data is being repackaged and enhanced by private sector providers, leading to the emergence of "big data" to support urban transportation planning efforts. The aging infrastructure will undergo further strain as agencies seek to implement technologies to support connected and automated vehicles and establish data governance to aid in the massive amounts of data anticipated to sustain and support these new operating systems. At the same time, agencies are capturing data regarding non-motorized and non-auto travel to support a variety of planning, health, and policy related initiatives. Agencies are also exploring open data initiatives and public-private partnerships to leverage the value of the data generated by the system. Beyond monitoring activities to private transport, the diffusion of new services such as shared

mobility and carpooling is generating additional amounts of data that can be exploited for better understanding urban mobility.

- On the demand side, researchers have observed changing travel patterns associated with changing demographics, emerging technologies, evolving land use patterns, and changing economic conditions. Transportation officials in urban areas are seeking to better understand these trends in order to improve the long-range planning processes and identify appropriate policy measures. Travel demand modelers are wrestling with methods to better capture and forecast these changes, as well as leverage big data now becoming available. The cost of collecting data using traditional means (census, travel survey programs) is skyrocketing, while research programs to investigate how emerging technologies could help support those programs are under-funded and slow in producing timely guidance. Crowdsourcing to have an instant feedback of travel conditions and gamification of on-the-fly travel surveys through apps (possibly coupled with GPS traces from the same device) are two examples of data sources that will probably play a role in the near future.
- The information systems currently in use by urban transportation agencies were not developed to meet demands for responsiveness and flexibility. The strain of the volumes of data being generated needs to be addressed, but within the context of the yet unknowns associated with the forthcoming automated and connected vehicle systems.

The transportation-related challenges and opportunities facing urban areas are not new. According to Barkley, the first traffic count programs can be traced back to 1844 in France, with the first US count made in 1885.

> *Most of the early counts were conducted in the cities, for one or more of the following reasons: (1) as a guide to the selection of suitable pavement surfaces, (2) to determine the causes of congestion and means of elimination, (3) to evaluate the effect of traffic on street cleaning, and (4) as an aid in the establishment of regulatory methods. (p. 3).*

During the mid-1920s, research began in earnest to estimate traffic when count data was not available and to predict usage of toll bridges (Barkley p. 4). Within twenty years, the field of urban data and information systems was firmly established, when, according to Barkley, "urban traffic planning, and particularly origin-destination surveys, has been given impetus by the Federal-Aid Act of 1944. This act made available $125,000,000 per year for the Federal-Aid Highway System in urban areas" (page 2).

Today, members of ABJ30 lead the research community in many of the following areas:

- Using census data in transportation planning;
- Emerging technologies and techniques to support travel time, speed, and reliability performance monitoring;
- Sources, uses and best practices with respect to emerging big urban data sources;
- Improving documentation of travel for non-motorized modes such as walk and bike;
- Understanding the intersection of health and transportation;
- Measuring and documenting demand for transportation through surveys and passive data;
- Researching the integration of all the above into systems to support Smart Cities and Megaregions.

In 2013, the American Association of State Highway Transportation Officials (AASHTO) developed Core Data Principles, which provide the framework for the activities pursued by ABJ30. These include:

1. Principle 1 – VALUABLE: **Data is an asset**—Data is a core business asset that has value and is managed accordingly.

2. Principle 2 – AVAILABLE: **Data is open, accessible, transparent and shared** —Access to data is critical to performing duties and functions, data must be open and usable for diverse applications and open to all.
3. Principle 3 – RELIABLE: **Data quality and extent is fit for a variety of applications**— Data quality is acceptable and meets the needs for which it is intended.
4. Principle 4 – AUTHORIZED: **Data is secure and compliant with regulations**—Data is trustworthy and is safeguarded from unauthorized access, whether malicious, fraudulent or erroneous
5. Principle 5 CLEAR: **There is a common vocabulary and data definition** —Data dictionaries are developed and metadata established to maximize consistency and transparency of data across systems.
6. Principle 6 – EFFICIENT: **Data is not duplicated** —Data is collected once and used many times for many purposes.
7. Principle 7 – ACCOUNTABLE: **Decisions maximize the benefit of data** Timely, relevant, high quality data are essential to maximize the utility of data for decision making.

PROCESS OVERVIEW

In order to get a historical perspective, we first read and reviewed the original Future of Urban Data that was written by this committee nearly 20 years ago. The thought was that doing so would give us an appreciation for what the community saw as the main topics that were going to be a focus in the opening decade of the 21st century before we sat to write about what we believe the next few decades hold for urban data.

The topics this paper covers were obtained using suggestions made by ABJ30's sub-committees. After these were received via email in late November of 2018 and compiled by the authors, this initial list was then used as a starting point for a brain-storming conference call in mid-December of 2018. Following this, the authors put together a draft of the paper using their personal and professional experiences in the selected topics then circulated the paper among the subcommittees for feedback and enrichment with technical references. Additional input was gathered during the committee's 2019 annual meeting at TRB.

We hope that the end result will provide a good representation of what the ABJ30 community believes will be some of the key topics in The Future of Urban Data as we approach the second decade of the 21st century.

LOOKING BACK WITH TODAY'S PERSPECTIVE

Twenty years ago, ABJ30 committee members conducted a future visioning exercise, the results of which anticipated that technology was expected to continue driving forward and providing new opportunities for urban data researchers while, at the same time, posing challenging data privacy and confidentiality issues. The same remains true today.

A notable change over the past twenty years has been the shift towards private companies becoming major urban data sources, whereas in the past government agencies were the main actors in collecting, compiling, archiving, and releasing urban data. This shift has come as a side-effect of our increasingly connected world, where we are surrounded by devices that are constantly monitoring the environment around them, tracking their movements, and communicating with cloud-based systems. The ever-decreasing costs of computing and storage have made it feasible to aggregate, process, and repackage incredibly large datasets.

Another marked change over the past two decades has been the rise of open source software, which has led to the rapid proliferation of viable alternatives to commercial data analysis tools for data processing, statistical analysis and modeling, simulation, machine learning, and visualization. The combination of accessible data and open source tools also helped grow and establish the new data scientist profession.

LOOKING FORWARD

After taking in information from the community, we believe that private source urban data and associated privacy issues will dominate the work and research in the area of urban data and information systems for the next decade. We also see the need for open data standards, documented data processing methods, and open source tools as playing a large role in this new era. The two following sub-sections cover the main topics contained in these two themes.

Private Industry-Controlled Data Sources

We are watching a transition from a world of government controlled and funded data collection, processing, and aggregation of urban data towards one where private-industry will be collecting, processing, packaging, and selling data in a continuous way. Whereas in the old paradigm government agencies would dictate what gets measured, where, and when; the new paradigm will be one where data is available for almost anywhere, but often collected and measured in different ways.

The spread of fast mobile data networks and advancements in electronics and battery technology was what made possible the rise of smartphones over the past decade. These are powerful and sensor-filled devices that are always on and connected to the internet. Passive movement data is a byproduct of the modern connected world.

A decade ago, the majority of mobile-derived movement data came from network activity (where phones had to switch from cell-to-cell); however, that data was known to be both too infrequent and low in locational accuracy. Nowadays, passive movement data is coming from smartphones running apps that use location as part of their functionality, including both foreground (while the app has focus) and background (while the phone is not in use or while some other app has focus) location collection. This category of data sources is often referred to using the acronym LBS for location-based services (Valentino-DeVries et al). With travel time probe data, providers are aggregating data from GPS-fleet management systems as well as smartphone data.

When combined, all these logged locations provide enough information for reconstructing an individual's travel patterns. Specialized companies known as aggregators then collect all these location pings and after some level of data cleaning and processing produce aggregations that are available for purchase by public agencies and their consultants. Some aggregators also sell individual device trajectory data. A large number of technology start-ups have moved into this space to create, package and sell data products to government agencies. In the process, these firms are now taking up space traditionally occupied by traditional planning, engineering and architectural consulting firms.

The location-based data has been phrased as "deep but narrow" referring to the fact that there could be millions of observations about origins and destinations of travel but little to no information available about the traveler, trip purpose, or travel mode. The more traditional behavioral data was collected by surveys, which were "shallow but broad" in that there were limited number of samples (sometimes 1% of the population) but broad in terms of the descriptors associated with the traveler, the trip purpose, the mode, etc.

In this rapidly approaching future, governmentally-sourced data may not survive and may go away unless steps are taken to protect the knowledge accumulated from running these collection programs over the past several decades and the details this data provides about the traveler and reasons for travel. The community believes that public agencies will need to step in and ensure survivability of data programs by actively engaging private industry through partnerships and other means. For example, although the emergence of ride sourcing services a few years ago, took local governments by surprise the recent emergence of electric scooters has been managed much more closely by cities across the United States. One notable example was regulation and permitting done by the city of Santa Monica, California which included a clause that all operational data from the connected scooters was to be made available to the government.

Even though private industry has the potential to overtake public data collection initiatives, we believe that government agencies will benefit from retaining some of their in-house data collection initiatives. These iniatives could serve as both benchmarks against which private data sources can be validated, provide a wealth of traveler information to fuse with, enhance the private sector data, and serve as expansion targets for privately-aggregated data. In the future, if agencies partner with emerging urban data providers, this could ensure that data sources, processing methods, aggregation, and expansion procedures result in proper representation in the final data products. Data fusion methods will be a growing area of research as they will be necessary to merge the traditional public-sourced small data with emerging private-sourced big data. Furthermore, improved data fusion methods will also ensure that the community will make the most out of emerging and traditional urban data sources. Government agencies may also have a role to play in the collection of data not deemed profitable for private agencies but determined to be critical to support transportation programs and initiatives.

Bicycle and pedestrian travel have historically been data deserts in transportation planning. The community focus over the past several years was to increase the availability of government-sourced count data. This was mainly achieved by installing counters and publishing best practices on how to process and aggregate counter data. At the same time, private data providers' initial focus was on vehicular traffic speed data (i.e., where is congestion?), with more recent offerings including origin-destination data, and even link-level flow data, with bicyclists and pedestrians appearing as separate categories (Schewel). Since this market segment is likely to remain small when compared to demand for vehicular travel data, it will be important for the public sector to remain engaged in order to ensure an equitable representation of these minor modes in emerging data products.

Equity will also be important when measuring and reporting on other less-used travel modes, such as transit. Although most passive transit data nowadays comes from system infrastructure (e.g., automated passenger counters, and automated fare collection and automated vehicle location), private companies behind emerging LBS data products are now starting to look at how to categorize transit travel that is already present in the location data streams that they process. Research by Graehler et al. has shown that ride-sourcing services compete with transit systems, but operational data from the current leading service providers (e.g., Uber and Lyft) is not always available to local transit operating and transportation planning authorities. That is mainly because ride-sourcing providers consider operational data as part of their intellectual property and thus heavily guarded. But that has the risk of making it so these same companies can indirectly set the research agenda by selectively providing access to their data.

Data Privacy

The abundance of data that leads to the emergence of private urban data sources described in the previous section has also led to less-than-careful system implementation and data distribution arrangements. In response to this loose handling of private and sensitive data, some governments have started rolling out laws and regulations, with the European Union's General Data Protection Regulation (GDPR) being the most notable example. The GDPR mandates that consent be obtained and made clear as to what the purpose of the collected data will be and also specifies that end users be able to remove consent at any point.

Recent events include the various data leaks and vulnerabilities reported by Facebook, inappropriate reselling of customer location data originated in US mobile operators (i.e., AT&T, T-MOBILE and SPRINT) (Whittaker), and a recent New York Times article (Valentino-DeVries et al.) that showed how much could be gleaned by looking at these data despite the fact that providers often qualify them as being anonymized. Events like these have the potential of impacting how much smartphone users will trust apps when they request access to location information in the future.

Historically, the Urban Data Committee focused on data that included a geographic component (i.e., origin-destination data). However, growing concerns around data privacy and accurate location data may impact the availability of these data. As such, a potential area for growing research will be that of methods that can protect data privacy while making available data that is detailed enough to support travel demand and mobility analyses. There is also the possibility that the availability of these data will go back to what it was decades ago before the advent of computerized geocoding and the global positioning system.

For example, LBS trace data largely exists because Google and Apple allow apps running on their mobile platforms to collect location data in the background after initial permission is granted. GPS data comes from smartphones, in-vehicle systems, and fleet management systems. This was not always the case and is not necessarily going to always be a feature of these platforms. It is also possible that smartphone users will stop allowing apps to collect location data, which must be explicitly authorized by users when first activating an app.

Finally, agencies will need to form new relationships with private data providers and become better educated about data end user agreements. In the past, agencies would own the data since they were responsible for funding its collection, processing and aggregation. Different from traditional data collection efforts, acquiring data from private sources does not necessarily guarantee that the public may get access to that data in the future. Agencies will need to carefully weight pros and cons of different data agreements to ensure that their purchase can meet both their internal needs as well as known external usages. Another area for future research will be on the acceptance testing and validation of data products as well as better measures and standards of accuracy.

Open Data Standards, Methods and Open Source Tools

The past two decades saw an exponential increase in the availability of urban data. The internet made it fast and easy to download data for analysis while open source tools have greatly increased the accessibility of sophisticated data analysis and visualization methods. These factors combined led to the rise of reproducible research where all data and code needed to recreate analyses are included with published research.

While initial open data efforts centered around making government-collected data available for download and use, more recent efforts have included the addition of application

programming interfaces (APIs) which allow data to be discovered, documented, and queried remotely in a programmatic way. We expect this trend to continue over the coming decades as internet connectivity and data download speeds continue to increase.

The US Chief Information Officer (CIO) Council (https://github.com/ombegov) has established a project on GitHub which promotes open data. The project also enumerates the following seven principles which open data efforts should be consistent with:

- *Public*. Consistent with the Office and Management and Budget's (OMB's) Open Government Directive, agencies must adopt a presumption in favor of openness to the extent permitted by law and subject to privacy, confidentiality, security, or other valid restrictions.
- *Accessible*. Open data are made available in convenient, modifiable, and open formats that can be retrieved, downloaded, indexed, and searched. Formats should be machine-readable (i.e., data are reasonably structured to allow automated processing). Open data structures do not discriminate against any person or group of persons and should be made available to the widest range of users for the widest range of purposes, often by providing the data in multiple formats for consumption. To the extent permitted by law, these formats should be non-proprietary, publicly available, and no restrictions should be placed upon their use.
- *Described*. Open data are described fully so that consumers of the data have sufficient information to understand their strengths, weaknesses, analytical limitations, security requirements, as well as how to process them. This involves the use of robust, granular metadata (i.e., fields or elements that describe data), thorough documentation of data elements, data dictionaries, and, if applicable, additional descriptions of the purpose of the collection, the population of interest, the characteristics of the sample, and the method of data collection.
- *Reusable*. Open data are made available under an open license that places no restrictions on their use.
- *Complete*. Open data are published in primary forms (i.e., as collected at the source), with the finest possible level of granularity that is practicable and permitted by law and other requirements. Derived or aggregate open data should also be published but must reference the primary data.
- *Timely*. Open data are made available as quickly as necessary to preserve the value of the data. Frequency of release should account for key audiences and downstream needs.
- *Managed Post-Release*. A point of contact must be designated to assist with data use and to respond to complaints about adherence to these open data requirements.

The increased availability of data has created new opportunities for research such as standardizing variable definitions and nomenclature to maximize data utility potential; developing data structures which can be shared by multiple applications and end users; greater understanding of data expansion methodologies to better understand data limitations; and data fusion techniques that add value to autonomous data through the combining of emerging and existing data sources to create a richer data that has both a higher number of observations and useful demographics.

Similarly, future research designed to better inform both researchers and practitioners on the above mentioned open principles would be beneficial to making the most out of what the future has to offer in urban data.

- Open source software tools
- The role of distributed repository platforms like GitHub

- Interacting with organizations like Zephyr foundation
- Standardize nomenclature and data structures
- Switch to using Application Program Interfaces (APIs) and software development and development operation methods (devOps)

A challenge for the urban transportation data community is to bridge the licensing and data use restrictions imposed by the private sector data providers (often resulting from their data sources and business practices) with the desire by agencies to provide the data in an open-source venue to their constituents, stakeholders, and consultants. In addition, the urban transportation data community is beginning to look beyond traditional training to consider data science, visualization, and software tools in order to not only address current needs but also leverage the opportunities afforded by these new data sources to create new applications and be ready to address emerging trends and technologies.

FINAL REMARKS

The future of urban data is a rich and bright one. The sheer volume and richness of data that we as a community are about to experience is likely to change how transportation policies and programs are tracked and evaluated. At the same time, the uncertainty about the future and longevity of these data sources will need to be addressed by ensuring that knowledge on traditional data collection and processing methods is not lost and can be properly converted into oversight and regulation by public agencies. The community should also be particularly concerned with issues surrounding equitable representation of all segments of the traveling public, as well as methods for continuing to make available data that may not be covered by private agencies due to low benefit-cost ratios perceived by these agencies.

REFERENCES

1. American Association of State Highway Transportation Officials (AASHTO). Core Data Principles. (nd). Available online at https://data.transportation.org/aashto-core-data-principles/.
2. Barkley, Robert Emmanuel. *Bibliography No. 11: Origin-Destination Surveys and Traffic Volume Studies*: Copyright, National Academy of Sciences, Washington, D.C., 1951. Reproduced with permission of the Transportation Research Board. Available online at www.travelsurveymanual.org/_attach/1.0/ca87d461a808fda4fbf7dc164d337abf6b4d42cc54b2ad6b/HRB-Biblio11-ODSurveysandTrafficVolumeStudies.pdf (last accessed January 31, 2019).
3. Graehler, M., R.A. Mucci, G.D. Erhardt. "Understanding the Recent Transit Ridership Decline in Major US Cities: Service Cuts or Emerging Modes?" Paper presented at the 2019 TRB Annual Meeting. Available online at http://usa.streetsblog.org/wp-content/uploads/sites/5/2019/01/19-04931-Transit-Trends.pdf (last accessed February 4, 2019)/
4. Schewel, Laura. "Applying LBS Data for Bike and Pedestrian Trip Estimation." Presentation given to the TRB Committee on Travel Survey Methods at the 2019 TRB Annual Meeting.
5. TRB Committee on Urban Data and Information Systems (ABJ30) Triennial Strategic Plan 2018-2020, revised January 2018.
6. Valentino-DeAVries, J., N. Singer, M.H. Keller, and A. Krolik. "Your Apps Know Where You Were Last Night, and They're Not Keeping It Secret." New York Times Interactive,

December 10, 2018. Available online at https://www.nytimes.com/interactive/2018/12/10/business/location-data-privacy-apps.html (last accessed February 4, 2019).
7. Whittaker, Zach. "Despite promises to stop, US cell carriers are still selling your real-time phone location data." Tech Crunch. January 9, 2019. Available online at https://techcrunch.com/2019/01/09/us-cell-carriers-still-selling-your-location-data/