PROJECT PROPOSAL: 2021-2022

Title: Mobility Analysis Workflow (MAW): An Accessible, Interoperable, and Reproducible Container System for Processing Raw Mobile Data

Principal Investigator: Cynthia Chen, Professor, Department of Civil and Environmental Engineering, University of Washington

1. Introduction/Problem Statement

The past two decades have seen a surge of studies using data from mobile devices to analyze individuals' mobility patterns (Chen et al. 2016). Such data often contains a large number of individuals (from hundreds of thousands to millions) while covering a sustained time period (from weeks to months and years). This data has two key pieces of information: the geographical locations (often expressed in longitude and latitude) where individual mobile devices are observed on the network, and the associated time when they are observed. Based on these two pieces of information, individuals' activity and travel patterns can be inferred.

Unlike travel surveys where trips are self-reported by the respondents and thus are automatically identified, data from mobile devices (hereafter "**mobile data**") needs to be pre-processed and analyzed to infer trips and their related information. This is due to the fact that mobile data is generated from users' opting into using certain mobile services (e.g., social media, or mobile apps) and consequently the amount and the quality of the mobile data vary greatly from one user to another depending on his/her usage patterns and the positioning technologies (e.g. GPS, WIFI, cellular tower, etc.) used by the data vendor. This means that pre-processing of the raw mobile data is first needed to correct data issues or to simply select a subset of users that meet certain criteria, followed by analysis of the pre-processed mobile data to infer mobility patterns. Those who use such data must go through these pre-processing and then analysis steps.

Exactly how this pre-processing and analysis sequence (called "**mobility analysis method**" hereafter for simplicity) is carried out is specific to individual researchers and indeed the proliferation of many studies that use big mobile data to infer mobility patterns suggest a variety of mobility analysis methods are being used (Chen et al. 2016). The different methods used can result in very different mobility patterns and those resulting mobility patterns can directly impact policy making. This motivates the first aim of the current study, the development of a Mobility Analysis Workflow (**MAW**) that incorporates pre-processing and analyzing the raw mobile data to infer mobility patterns, which everyone can access and use with their own mobile data. More specifically, MAW supports four important properties of open science: accessibility, interoperability, reproducibility, and reusability. Accessibility requires not only the sharing of the code used for mobility analysis but also providing an easy-to-install and easy-to-configure version of the computer program or software with necessary dependencies, documentation, and licenses. In the context of mobility analysis, interoperability means a mobility analysis workflow can be tested and executed on different computing and operating systems. Reproducibility is achieved if the same results are obtained when a mobility analysis workflow is tested with the same input data (Benureau and Rougier 2018) by a different researcher or on a different computing system. Reusability implies that a workflow can be reused over additional cases other than those for which it was originally designed for (Lamprecht et al. 2020).

In transportation research and practice, the use of accessible, interoperable, reproducible, and reusable workflows is still at a nascent stage. Existing efforts focus on the distribution of developed programs (typically in the form of code scripts) through platforms such as GitHub, JupyterLab or Google Colab (Boeing 2020; Majka et al. 2019). As noted earlier, simply making scripts open-source does not address the dependency issues as aforementioned. To run the scripts, users need to figure out the compatible types of machines, operating systems, script compilers, and versions of packages and libraries. This information is not always clearly defined in a code repository, making it challenging to reproduce the research work and results. Though containers have recently attracted transportation researchers' attention (Feygin et al. 2020), they are used primarily to allow users to run models remotely, rather than for designing and implementing workflows.

2. Project Objectives

Even though no two mobility researchers use exactly the same methods to process and analyze mobile data, there are main commonalities and differences. In the pre-processing stage, as noted in Wang and Chen (2018), presence of oscillation is a primary issue in mobile data and not removing it can result in an overestimation of the regularity commonly found in human mobility patterns. Yet, there exist a variety of available methods and parameter settings in correcting oscillations, even for the same dataset. In the data analysis stage to derive the mobility patterns, even though clustering is a common method to identify stays (places where people perform activities), different clustering algorithms have been used (e.g., trace segmentation vs incremental clustering) with different threshold values (e.g., maximum distance or minimum duration required). Currently there is little to no systematic knowledge regarding how the differences in pre-processing and analysis of the mobile data impact the derived mobility patterns. Answering this question comprises the second aim of the study. To answer this question, six workflows that employ different mobility data pre-processing and analysis methods are designed to pre-process and analyze two commonly-used sample mobile datasets, and the resulting mobility patterns are compared among the six workflows.

3. Proposed Methodology and Data

To test how the inferred mobility patterns are affected by different pre-processing or analysis algorithms, their respective orders in a workflow, and change point values, two sample datasets will be drawn from the real-world app-based data and tested in different workflows. One dataset consists of predominantly GPS location records and the other of predominantly cellular location records. Additionally, six workflows are designed (see Table 1) to test the uses of various pre- and processing methods on the resulting mobility metrics.

Table 1. Design of Six Workflows to test the uses of pre- and processing methods on resulting mobility metrics

Workflow index	*Workflow design*	*Note*
1	Incremental Clustering – Stay Duration Calculator	Oscillation is not dealt with.
2	Incremental Clustering – Stay Duration Calculator – Oscillation Corrector – Stay Duration Calculator	Oscillation is addressed as a post-processing step.
3	Oscillation Corrector – Incremental Clustering – Stay Duration Calculator	Oscillation is addressed as a pre-processing step.
4	Incremental Clustering – Stay Duration Calculator	Only incremental clustering is applied.
5	Trace Segmentation Clustering – Stay Duration Calculator	Only trace segmentation clustering is applied.
6	Trace Segmentation Clustering – Incremental Clustering – Stay Duration Calculator	Both trace segmentation clustering and incremental clustering are applied.

Two sets of metrics will be evaluated, relating to the inferred mobility patterns and the computational costs. The first category include: number of trips per person per day, radius of gyration and departure time distribution. The second category of performance metrics relates to the computational performance of workflows. The key metrics in this category are the computation time of a given workflow, measured as the time interval from when the workflow begins reading input data to when it finishes printing the output, and the memory usage over time throughout the running of the workflow.

4. Work Plan (Project Tasks)

Task 1. Literature review. We will review two parts of the literature, the first part relating to accessibility, interoperability, reproducibility and reusability and the second part relating to the pre-processing algorithms and analysis methods used in deriving mobility patterns from the big data. We expect most of the literature review for the first part will be in computer science area, as in transportation research and practice, existing efforts focus on the distribution of developed programs (typically in the form of code scripts) through platforms such as GitHub, JupyterLab or Google Colab (Boeing 2020; Majka et al. 2019). For the second part, a comprehensive literature review will be carried out, as the inferring mobility patterns from big data has been done in a variety of disciplines including transportation, urban planning, geography, social science and computer science.

Task 2. Development of Mobility Analysis Workflow (MAW)

Subtask 2.1. Container Design.

Based on the existing literature on the various pre-processing and analysis methods used to infer mobility patterns from big data, we plan to develop five constainers: trace segmentation clustering, incremental clustering, stay duration calculator, oscillation corrector and stay integrator. We will review suitability of these five containers based on the result of Task 1.

Subtask 2.2. Contain development.

Table 2. Initial design of each of the five containers to be developed.

Container name	**Input and Output**
Trace Segmentation Clustering	*Input*: Location records[1] on a single day, sorted by time. *Output*: Same as inputs, with identified stay locations[2] added. Transient points[3] are denoted as -1.
Incremental Clustering	*Input*: Location records (on either one day or multiple days). *Output*: Same as inputs, with identified stay locations added. Transient points are noted with placeholder of -1.
Stay Duration Calculator	*Input*: Location records with stay locations identified. *Output*: Same as input, with stay durations added. Transient points have a placeholder (e.g., -1) for stay durations.
Oscillation Corrector	*Input*: Location records with or without stay information, sorted by time. *Output*: Same as inputs with location records caused by oscillations removed.
Stay Integrator	*Input*: Two sets of location records with inferred stay information attached, from two types of data (e.g. GPS and cellular), respectively. *Output*: Location records with stay information (stay location and duration) attached.

Subtask 2.3. Workflow designs. Table 3 provides our plan for the development of six workflows. **Workflows 1 through 3** address the situations where oscillation is simply ignored, dealt with in post-processing (i.e. after inferring stays), and corrected in pre-processing (i.e. before inferring stays), respectively. Different change point values of Oscillation Corrector are tested in workflows that include this container. In both workflows 1 and 2, incremental clustering is applied directly to the raw cellular data, followed by one step to calculate the durations of inferred stays. Workflow 2 additionally post-processes the inferred stays by removing those suspected for oscillation. Since this operation may change the cluster compositions (e.g. two clusters merged into one), a Stay Duration Calculator container is subsequently called to update the stay durations. Workflow 3 in contrast, addresses oscillation beforehand, prior to applying incremental clustering and calculating stay duration. **Workflows 4, 5** and **6** concern the situations where only incremental clustering, only trace segmentation clustering, and both trace segmentation clustering and incremental clustering are applied to the GPS data, respectively. In each workflow, a Stay Duration Calculator container follows the clustering step(s) to calculate the durations of inferred stays. For workflow 6, the particular order in which trace segmentation clustering comes ahead of incremental

clustering is based on the consideration that when the two clustering methods are applied together, the purpose of incremental clustering is to identify recurring stays over multiple days. Therefore, incremental clustering needs preliminary stays inferred by trace segmentation clustering as its inputs.

Table 3. Workflow Designs

Workflow index	*Workflow design*	*Note*
1	Incremental Clustering – Stay Duration Calculator	Oscillation is not dealt with.
2	Incremental Clustering – Stay Duration Calculator – Oscillation Corrector – Stay Duration Calculator	Oscillation is addressed as a post-processing step.
3	Oscillation Corrector – Incremental Clustering – Stay Duration Calculator	Oscillation is addressed as a pre-processing step.
4	Incremental Clustering – Stay Duration Calculator	Only incremental clustering is applied.
5	Trace Segmentation Clustering – Stay Duration Calculator	Only trace segmentation clustering is applied.
6	Trace Segmentation Clustering – Incremental Clustering – Stay Duration Calculator	Both trace segmentation clustering and incremental clustering are applied.

Subtask 2.4: implementation of developed workflows on actual data. Once the workflows are developed, they will be applied to the dataset used for this project. The data used in this study is the GPS and cellular data generated when people use mobile phone apps (thus called app-based data) in the central Puget Sound region covering the four counties (King, Kitsap, Pierce, and Snohomish) during the period of March, April, and November of 2019. We then randomly selected 1,000 users whose location records contain over 80% GPS location records. This dataset constitutes our sample GPS dataset. Similarly, a random set of 1000 users each with more than 80% of location records as cellular location records are selected, which comprises the sample cellular dataset.

Task 3. Analysis of results and interpretation. Results from Task 2 above will be analyzed. As noted earlier, the effects of the various pre-processing and analysis algorithms on the big data for deriving mobility patterns will be assessed with two kinds of performance metrics: mobility related metrics and computational metrics.

Task 4. Final report will be completed encompassing Tasks 1 to 3.

5. Project Schedule (1/2 – 1 page)

Table 4 below lays out the proposed project tasks.

Table 4. Proposed Work Plan for the Project

Project Tasks	**Task Description**	**Task Duration**	**Milestone**
Task 1: Literature review	Two parts of review will be conducted: the first part relating to general and transportation-related literature on accessibility, interoperability, reproducibility and reusability (primarily on computer science); the second part relating to the algorithms used in deriving mobility patterns from the big data	1 month	A literature review synthesis
Task 2: Development of Mobility Analysis Workflow (MAW)	Designs of containers that would encompass all essential pre-processing and analysis steps on the big data to derive mobility patterns	8 months in total	Developed workflows in open-source github and docker container repository
Subtask 2.1: container design	Each container shall correspond to a stand-alone task frequently used in pre-processing or analysis of the big data for deriving mobility patterns	1 month	
Subtask 2.2: container development and testing	Once the containers are designed in subtask 2.1, work will be carried out to actually develop the containers and test them	3 months	
Subtask 2.3: workflow designs	Once containers are developed and tested, work will be carried out to develop workflows that are a sequence of containers.	1 month	
Subtask 2.4: implementation of developed workflows on actual data	Apply the developed workflows on actual data and obtain statistics on performance metrics, both relating to the mobility patterns and the computational performance	3 months	
Task 3: Analysis results and interpretation	Analyze the results on the results of subtask 2.4.	2 months	Analysis results
Task 4: Report completion and dissemination	Complete the final report and disseminate the results	1 month	Final report

6. Relevance to the Center Theme/Mission

Mobility analysis, or understanding and modeling of people's mobility patterns in terms of when, where, and how people move from one place to another is fundamentally important. Such information is not only important for answering many scientific inquiries regarding how people interact with urban spaces and with each other, but also as a basis for many large- or mega-scale investment decisions on the nation's multi-modal transportation infrastructure. For decades, information on people's mobility patterns has been obtained from self-reported household travel surveys where randomly-selected respondents are asked to report all of their travel on one or two pre-determined travel survey days (Stopher and Greaves 2007). Travel surveys, though providing rich information, are expensive (about $250-350 per household), and have relatively small sample sizes (typically ~0.1% of the region's population for urbanized areas). Because household travel surveys are conducted rather infrequently (once every few years), they are unsuitable for answering questions relating to how mobility patterns evolve over time or change after events.

The past two decades have seen a surge of studies using data from mobile devices to analyze individuals' mobility patterns (Chen et al. 2016). Unlike travel surveys where trips are self-reported by the respondents and thus are automatically identified, data from mobile devices (hereafter "**mobile data**") needs to be pre-processed and analyzed to infer trips and their related information. This is due to the fact that mobile data is generated from users' opting into using certain mobile services (e.g., phone services, social media, or mobile apps) and consequently the amount and the quality of the mobile data vary greatly from one user to another depending on his/her usage patterns and the positioning technologies (e.g. GPS, WIFI, cellular tower, etc.) used by the data vendor. This means that pre-processing of the raw mobile data is first needed to correct data issues or to simply select a subset of users that meet certain criteria, followed by analysis of the pre-processed mobile data to infer mobility patterns. Those who use such data must go through these pre-processing and then analysis steps.

Unfortunately, there is no standard or even guidelines in terms of what kinds of pre-processing and analysis methods shall be applied to what kinds of data and what types duration and distance thresholds need to be used. The proposed project is to address this gap by developing a set of accessible, interoperable, reproducible and reusable dockers that anybody can use on their own data and answer the questions above. As the field is moving toward using more big data for deriving travel patterns, it is critical that we as a field obtain sufficient understanding about how the use of our own methods can have an effect on the resulting mobility patterns.

7. Anticipated Outcomes and Deliverables

Anticipated outcomes and deliverables include a final report that documents all research conducted in the proposed project as well as a set of docker containers on github for everybody to use on their own big data to derive mobility patterns (see Table 4).

8. Research Team and Management Plan

Team Expertise

Cynthia Chen is a professor in the Department of Civil & Environmental Engineering at the University of Washington (Seattle). She is an internationally renowned scholar in transportation science and directs the THINK (Transportation-Human Interaction and Network Knowledge) lab at the UW. The work at THINK lab is at the intersection of human behavior and the system within which individuals and businesses operate. More specifically, THINK lab research unpacks complexities found at different scales of an urban system from micro-level individual behaviors, to meso-scale interactions formed as the result of individual behaviors, and to macro-scale system behaviors that propagate through a single network or multiple networks. Cynthia has published over 60 peer-reviewed publications in leading journals in transportation and systems engineering including Transportation Research Part A-F and Omega, as well as interdisciplinary journals such as PNAS. Her research has been supported by federal agencies such as NSF, NIH, APAR-E, NIST, USDOT, and FHWA as well as state and regional agencies. Cynthia served a two-year assignment from 2017-19 as the Program Director of Civil Infrastructure Systems, CMMI (Civil, Mechanical, and Manufacturing Innovation) division with the National Science Foundation. She is an associate director of TOMNET (Center for Teaching Old Models New Tricks), a USDOT-funded Tier 1 University Transportation Center led by ASU. She is also an associate editor of two leading journals: Transportation Science, and Service Science and is on the editorial board of Sustainability Analytics and Modeling.

Ka Yee Yeung focuses on the development of machine learning tools, their application to computational biology and the development of containerized tools to enhance the reproducibility of research. My research focuses on the development of methods and containerized cloud-enabled software tools to facilitate the reproducible analyses of big biomedical data. I also develop machine learning methods that blend both computer science and statistics for applications in bioinformatics.

Wes Lloyd engages in research spanning Serverless computing (FaaS), Containerization, Infrastructure-as-a-Service (IaaS) cloud computing, virtualization, performance and cost modeling of cloud application deployments, and cloud infrastructure management. By harnessing resource utilization metrics to characterize performance of applications and workloads running in the cloud, we develop and apply machine learning and statistical techniques to model application performance, address scaling, and improve autonomous resource management. Given a variety of ways applications can be deployed to cloud environments, our research aims to demystify the myriad of options to guide scientists and practitioners towards making informed deployment decisions that improve performance and availability while reducing hosting costs.

Management and Communication plan.

The team plans to meet on a bi-weekly basis from start to end. Two students and a postdoc will be involved in the project team and they will participate in all meetings. Every quarter, an in-person meeting of all team members will be held either on the Seattle or Tacoma campus.

9. Technology Transfer Plan

The following technology transfer plan will be carried out:

- Final report and paper to be submitted to either arXiv.org or journals for publication;
- Tutorials developed for users who are interested in using the docker containers that are to be developed;
- Youtube video to illustrate the use of the developed docker containers;
- Presentations at conferences and workshops

10. Workforce Development and Outreach Plan

The proposed project will involve two graduate students and a postdoc from the University of Washington's Seattle and Tacoma campuses. In addition, we will recruit undergraduate students from both campuses to work on the project. Underrepresented and minority students are always a priority to be recruited.

11. References

Benureau, Fabien C. Y., and Nicolas P. Rougier. 2018. "Re-Run, Repeat, Reproduce, Reuse, Replicate: Transforming Code into Scientific Contributions." *Frontiers in Neuroinformatics* 11. https://doi.org/10.3389/fninf.2017.00069.

Boeing, Geoff. 2020. "The Right Tools for the Job: The Case for Spatial Science Tool-Building." *Transactions in GIS* 24 (5): 1299–1314. https://doi.org/10.1111/tgis.12678.

Chen, Cynthia, Jingtao Ma, Yusak Susilo, Yu Liu, and Menglin Wang. 2016. "The Promises of Big Data and Small Data for Travel Behavior (Aka Human Mobility) Analysis." *Transportation Research Part C: Emerging Technologies* 68 (July): 285–99. https://doi.org/10.1016/j.trc.2016.04.005.

Feygin, Sidney A., Jessica R. Lazarus, Edward H. Forscher, Valentine Golfier-Vetterli, Jonathan W. Lee, Abhishek Gupta, Rashid A. Waraich, Colin J. R. Sheppard, and Alexandre M. Bayen. 2020. "BISTRO: Berkeley Integrated System for Transportation Optimization." *ACM Transactions on Intelligent Systems and Technology* 11 (4): 38:1-38:27. https://doi.org/10.1145/3384344.

Lamprecht, Anna-Lena, Leyla Garcia, Mateusz Kuzak, Carlos Martinez, Ricardo Arcila, Eva Martin Del Pico, Victoria Dominguez Del Angel, et al. 2020. "Towards FAIR Principles For Research Software." *Data Science* 3 (1): 37–59. https://doi.org/10.3233/DS-190026.

Majka, Kevin, Eric Nagler, Alex James, Alan Blatt, John Pierowicz, Panagiotis Ch Anastasopoulos, and Grigorios Fountas. 2019. "Applications of Knowledge Discovery in Massive Transportation Data: The Development of a Transportation Research Informatics Platform (TRIP)." FHWA-HRT-19-008. Washington, DC: Federal Highway Administration. https://trid.trb.org/view/1577767.

Stopher, Peter R., and Stephen P. Greaves. 2007. "Household Travel Surveys: Where Are We Going?" *Transportation Research Part A: Policy and Practice*, Bridging Research and Practice: A Synthesis of Best Practices in Travel Demand Modeling, 41 (5): 367–81. https://doi.org/10.1016/j.tra.2006.09.005.

12. Qualifications of Investigators

Cynthia Chen

Department of Civil & Environmental Engineering
University of Washington
Box 352700, More Hall 133A
Seattle, WA 98195-1750

qzchen@uw.edu
phone: 206.543.8974
THINK LAB: https://sites.uw.edu/thinklab

I. PROFESSIONAL PREPARATION

Postdoctoral Fellowship, 5/2002–8/2003
Project Focus: Development of a mass-scrappage model for vehicle emissions
Advisor: Debbie Niemeier, University of California, Davis

Ph.D. Civil and Environmental Engineering, 1996-2001
Dissertation: An Exploration of Activity Scheduling and Rescheduling Processes
Advisor: Ryuichi Kitamura, University of California, Davis, CA

M.S. Transportation Engineering, 1993-1995
Thesis: A Comparison between Revealed Preference (RP) and Stated Preference (SP) based on Simulation Results
Advisor: Patrick Beaton, New Jersey Institute of Technology, Newark, NJ

B.A. Tourism, 1988-1992
Nan Kai University, Tianjin, China

II. KEY APPOINTMENTS

2016–present	Professor of Civil and Environmental Engineering, University of Washington
	Adjunct Professor of Urban Design and Planning, University of Washington
2017–2019	Program director of Civil Infrastructure Systems, CMMI (Civil, Mechanical, and Manufacturing Innovation) division, National Science Foundation
2009–2016	Associate Professor of Civil and Environmental Engineering, University of Washington
	Adjunct Associate Professor of Urban Design and Planning, University of Washington
2003–2009	Assistant Professor of Civil and Environmental Engineering, City College of New York

III. FIVE MOST RELEVANT PEER-REVIEWED PUBLICATIONS *(superscript denotes graduate students and postdocs from my group)*

1. Wang*, F.; Wang, J.; Cao*, J.; Chen, C.; and Ban, J. (2019) Extracting Trips from Multi-Sourced Data for Mobility Pattern Analysis: An App-Based Data Example. *Transportation Research Part C* 105, 183-202. https://doi.org/10.1016/j.trc.2019.05.028
2. Wang*, F. and Chen, C. (2018) On data processing required to derive mobility patterns from passively-generated mobile phone data. *Transportation Research Part C*, 87, 58-74. DOI: 10.1016/j.trc.2017.12.003.
3. Chen, C.; Ma, J.; Susilo, Y.; Liu, Y and Wang*, M. (2016) The promises of big data and small data for travel behavior (aka human mobility) analysis. Transportation Research Part C, 68, 285-299.
4. Chen, C.; Bian, L. and Ma, J. (2014) From traces to trajectories: how well can we guess activity locations from mobile phone traces? *Transportation Research Part C* 46, 326-337.
5. Mokhtarian, P. and Chen, C. (2004) TTB or Not TTB, that is the Question: A Review and Analysis of the Empirical Literature on Travel Time (and Money) Budgets. *Transportation Research Part A* 38(9-10), *643-675.*

13. Budget Including Non-Federal Matching Funds

Institution: University of Washington
Project Title: Mobility Analysis Workflow (MAW): An accessible, interoperable, and reproducible container system for processing raw mobile data
Principal Investigator: Cynthia Chen

Budget Period: 8/1/2021 - 07/31/2022

CATEGORY	Budgeted Amount from Federal Share	Budgeted Amount from Matching Funds	Explanatory Notes; Identify Source of Matching Funds
Faculty Salaries	30,445	7,530	2.00 summer months for Cynthia Chen. 0.45 academic months will be used as cost share for Cynthia Chen
Other Staff Salaries			
Student Salaries	60,803	36,689	2 student support. 1 student support will be used for cost share.
Fringe Benefits	20,927	10,010	24.0% for faculty; 22.4% for student
Total Salaries & Benefits			
Student Tuition Remission	43,995	50,589	4 quarters of support for each student on project. 3 academic quarters of non-resident tuition waivers for each student will be used as cost share.
Operating Services and Supplies			
Domestic Travel	1,012		Attend workshops organized by the center and conferences
Other Direct Costs (specify)			
Other Direct Costs (specify)			
Total Direct Costs	157,182	104,818	
F&A (Indirect) Costs	62,818	5,182	MTDC 55.5%
TOTAL COSTS	220,000	110,000	

Grant Deliverables and Reporting Requirements for UTC Grants (November 2016)

Exhibit F

UTC Project Information	
Project Title	
University	
Principal Investigator	
PI Contact Information	
Funding Source(s) and Amounts Provided (by each agency or organization)	
Total Project Cost	
Agency ID or Contract Number	
Start and End Dates	
Brief Description of Research Project	
Describe Implementation of Research Outcomes (or why not implemented) Place Any Photos Here	
Impacts/Benefits of Implementation (actual, not anticipated)	
Web Links • Reports • Project Website	