Proceedings of the 2023 Winter Simulation Conference
C. G. Corlu, S. R. Hunter, H. Lam, B. S. Onggo, J. Shortle, and B. Biller, eds.

Data-Driven Agent-Based Modeling to Improve Bus Route Effectiveness: A Case Study

Devon Zillmer
Anna Tucker
James K. Starling

Center for Data Analysis and Statistics
Department of Mathematical Sciences
United States Military Academy
606 Thayer Road
West Point, NY 10996, USA

Bryan Hartman
Paul Kunnas

Knowledge Management and
Assessments Division
Eighth United States Army
USAG Humphreys, Republic of Korea

Abstract

Located just outside of Pyeongtaek, Republic of Korea, Camp Humphreys is one of the largest United States overseas military bases in the world. In 2022, the installation identified that the bus system was not effectively used and directed a review of the routes and a proposal to restructure if necessary. Operations Researchers organized data collection of over 5,000 bus route observations, and the data were used to define essential characteristics of the route: oriented stop demand, probability of passengers departure by stop, and inter-stop travel time. Using this information, a simulation was built in Python to model the average buses on each route. Then new proposed bus routes were analyzed in the simulation to assess if they provided an improved average travel time. In simulation, the new routes offered a reduction of around 25% of the average travel time, or about 3 minutes for the average traveller.

1 Introduction

Located about 40 miles south of South, roughly 100 miles away from the Demilitarized Zone in the Republic of Korea, Camp Humphreys is one of the largest United States overseas military bases in the world. With over 30,000 Servicemembers residing on the installation, it has grown into a small city since the close of active hostilities in the Korean War. During the spring of 2022, the 403rd Army Field Support Battalion - Korea (403 AFSBn) requested assistance from the resident headquarters, the 8th U.S. Army (8A), to help improve the efficiency of the buses on post. The issue was that many Servicemembers were paying to use taxis to travel around on post, despite having over a dozen buses running full-time during the duty day.

There had been buses on post for several decades, but as the size of the installation increased, including new construction and additional units being housed on Camp Humphreys, the bus routes grew long and tenuous.

There are three routes operating on the installation (Red, Blue, and Green), visualized in Figure 1. Each of the routes consisted of a long loop of 35 or more stops, with each bus taking over an hour to run the complete loop. Rather than focusing a route on a targeted populations (e.g., from a major barracks cluser to the main shopping center, or from major workplaces to dining facilities), each bus route covered the entire installation, resulting in inefficient routes, frequent stops, and slow travel times.

These long travel times, with their associated instability of arrival times due to the long routes with many stops, combined with an inability to provide more buses at high demand times (e.g., the transition from morning physical activities to breakfast, or after close of business to dinner) due to limited staffing and buses, all contributed to the population on the installation having low confidence in the ability of the bus system to efficiently get travellers where they need to go. These factors increased the utilization of taxis

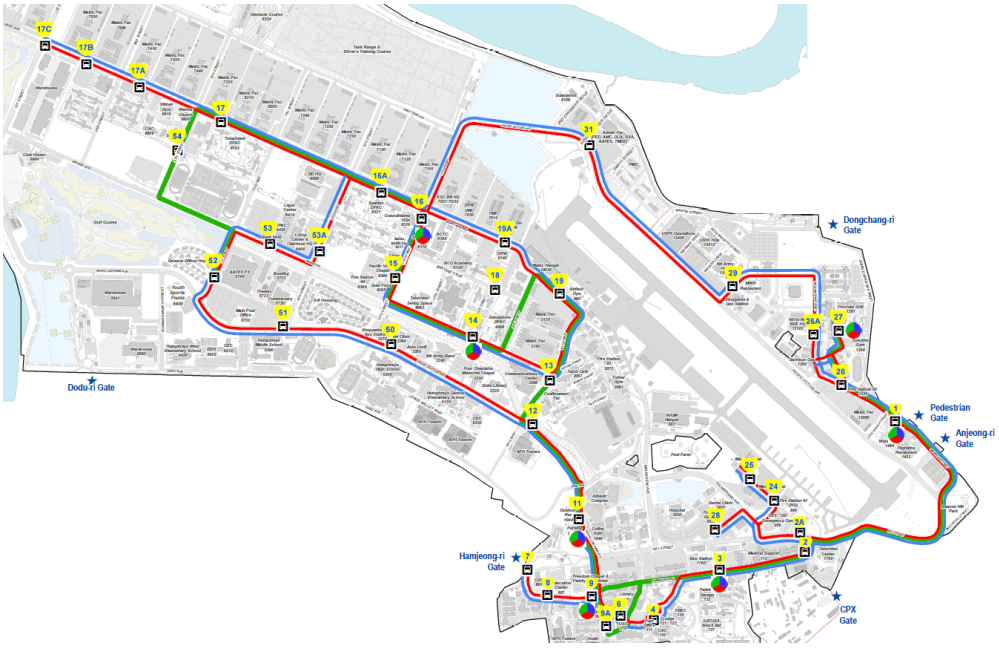

Figure 1: Graphic depicting three original bus routes.

on the installation. Further complicating efforts to improve the bus routes was the rudimentary technology of the system. Without systems such as automated data collection or built-in system to record bus arrival times or passenger flows, no systematic records of bus utilization existed outside of anecdotal observations. As a result of this, the 8A leadership was not in a position to be able to easily use data-driven analysis to make a decision on how to improve the system.

To begin to address the potential issue, the authors, serving as the 8A Operations Researcher and Systems Analysts (ORSAs), gleaned data from two sources. One source of data was a pull of the taxi log over a three-month period, to better understand the unmet needs of the bus transportation system. To gather data on bus use, a plan was developed to send observers to record bus utilization, varying the routes, boarding times, and boarding locations. These observations took place over the course of a month, covering 5,009 bus stop observations over roughly 160 hours, and resulted in 144 records of bus utilization.

The object of this endeavour was to provide a recommendation for new bus routes. To this end, analysis of which stops are significant, visualization to 8A leadership, creation of a functional simulation modeling the current bus ridership, development of a metric to assess efficacy of a route, and an analysis of proposed routes in the simulation to assess if the new routes would be an improvement.

2 Literature Review

There is a wide array of content on modeling public transport, ranging from simple tutorials on discrete event simulations using bus routes to sophisticated models oriented on optimizing assorted components of large city bus fleets.

Several researchers have developed sophisticated models to assess and optimize various aspects of the bus routes, including Vázquez-Abad and Fenn (2016), López et al. (2019), and Shi et al. (2021). In Vázquez-Abad and Fenn (2016), the author uses mixed optimization to consider optimal fleet size for an airport with a target wait time for the audience at the airport. While this problem is about bus optimization, it is focused on fleet size and stochastic processes related to passenger arrival, rather than adjustments to the route to improve route efficacy. López et al. (2019) also considers airport bus routes, but focusing on bus and fleet sizing for feasibility. Because of the focus of these papers, neither proved a major inspiration for

this paper's work. In Shi et al. (2021), the authors use very detailed data and the Analytic Hierarchy Process to build a comprehensive evaluation model for bus route optimization. While thorough, the authors' model relies on more complete data and more complex inputs, whereas this paper has much less complete data and has to make concrete recommendations. These types of analyses are comprehensive and sophisticated, but can be challenging to clearly explain decisions outside of a technical situation.

Pereira and Chwif (2018) provided a bus modeling simulation that was oriented on a larger urban area, focused on different metrics. In their work, a framework for modeling initial bus route behavior via simulation is proposed, very much inspiring the process used for this paper's analysis. From the modeling perspective, their work was comparable, but was conducted in SIMUL8 and made some different assumptions about the stochastic processes that can be used to model reality. Their analysis considered some modifications of routes, but was largely not applicable due to its focus on a large, urban collection of routes and the infeasibility of their conclusions (i.e., long waiting times, which works contrary to the motivation of the current paper's research goal). Ultimately, their approach did not fully translate to the current work, due to different requirements motivating the analysis.

Previous analysis on subway routes, as in Schmaranzer, Braune, and Doerner (2016), relate directly to the problem presented in this paper. The methods, including analysis of demand, discussion of hypothetical circumstances, and simulation to estimate effects of varying conditions, all directly inform the processes used here. While the situation, with its restrictions and metrics considered, are different from the current work, its approach inspired much of the work in the current paper.

The most relevant previous work was from Pemberton (2020), and provided many similar components of analysis as those desired by the authors. The paper provided analysis of transportation routes covering a city, and has several examples of visualization, clear definition of a metric for improving the bus routes, and discussion on how to restructure bus routes to improve efficacy. The framework used by Pemberton (2020) provided an excellent initial perspective to begin this paper's analysis. Because of the scope of the work, breadth of data, and considerations of coverage, many of the specific recommendations and considerations did not fully apply to the current work.

3 Methods

To explain the process used to provide recommendations, the following will outline the data collection, analysis of the data, development of a simulation model, and then implementation of new routes in the simulation.

3.1 Data Collection

There were two major data sets collected for this paper's analysis: taxi pickups and bus boarding/exiting. The taxi data were provided by Exchange Taxi Services, and consisted of a CSV record of calls made for taxi pickup, and included date, time, and location for pick-up (usually a building number or colloquial name) for every call on Camp Humphreys from 1 May to 31 Aug 2022, with approximately 422,000 entries. Of these total entries, only the most common entries were used, reducing the dataset to 358,204 entries.

The bus data were harder to collect. A small team of Korean Augmentees to the United States Army (KATUSA, who are Republic of Korea Citizens serving their obligatory service time as Korean Soldiers working with the U.S. Army) was assembled by the 8A ORSAs, and rode on 145 different buses between 17 Oct and 28 Oct 2022, including all three bus routes at assorted times during the duty day. Due to external challenges, they were not able to collect data on all possible times of day and days of the week. Each KATUSA rode the bus for a full loop, boarded and exited at the same place, and at each bus stop recorded the time, the count of passengers who boarded, and the count of passengers who exited the bus.

This data collection plan did not collect the most commonly-sought data type for public transit analysis: the pick-up and drop-off pairing of riders (e.g., this passenger boarded stop 2 and existed stop 45), did

not directly count the current number of passengers on the bus at point in time (i.e., to assess utilization directly), or a significant depth of collection by hour (i.e., only a few observations per route per hour).

At the conclusion of the data collection, the bus data were then entered into a CSV recording route, stop, date and time, and quantity picked up and dropped off. This produced some 5008 raw entries across the 145 observations.

3.2 Data Analysis

In order to propose a better route, some analysis of the data was required. Specifically, tools to visualize unmet demand (i.e., using taxi pickup data) and visualization of important stops (i.e., high demand versus unused stops), with additional requirements to assess bus utilization. Additionally, to facilitate simulation, three key sets of values had to be calculated: directional demand at each stop (or average number of pick-ups per directional stop), probability of exit at each stop (average percent of bus exiting at a directional stop), and interstop time (average travel time between each stop). The following discusses each of these five topics.

3.2.1 Taxi Data

The taxi data required significant assumptions to be useful. Because the data were hand-typed, there was significant variation in pick-up locations that were all the same (e.g., "p= 2908" compared to "p=2098 front") that had to be accounted for and binned to reduce the unique pickup locations and account for accurate pickup requests for a given building. Then, the 93 pickup locations with more than 1000 pickups in the data set were each given a meaningful label (i.e., Talon Dining Facility), and assigned a GPS coordinate approximated by lookup on an installation map and publicly available imagery. This data were encoded into a lookup dictionary and visualized in a Python Jupyter Notebook using the Seaborn package. This allowed for interactive products to facilitate Camp Humphreys Garrison leadership "see" the demand in space. The subsequent use of the data will continue later on.

3.2.2 Oriented Demand

Initial analysis of the bus stops considered the stops with the highest and lowest numbers of pick-ups and drop-offs, binned only by bus stop. As subsequent simulation revealed, this was incomplete, and further analysis had to be conducted to consider directional stops. For consistency, each stop on a route was classified as either "going towards the main gate" or "away from the gate". This was chosen rather than purely directional orientation (i.e., stop 3-North and 3-South) to contextualize the demand signal, make adding new stops easier, and to facilitate intuition when adding in the additional taxi demand signal.

3.2.3 Bus Data

Because the data collection were completed before any analysis could occur, some additional values had to be extrapolated from the bus observations. One example of this was an estimate of the current number of passengers on the bus. To do this, a "current bus load" was calculated by starting with number of pick-ups at the first observed stop (for each of the existing routes, the route was a continuous loop, not a route that began at a bus depot and ended there; as such the "first stop" was the first observed bus stop). Then for each subsequent stop, adding or subtracting the number of pick-ups or drop-offs as necessary. Once the current load was calculated for an entire route observation, if there was any negative current load, the magnitude of the largest negative number was used as the starting passenger count, and added to the initial bus load. This ensured the current load was always non-negative. The key output from the current bus load computation was the ability to visualize bus utilization over a route.

3.2.4 Bus Simulation Inputs

Three sets of numbers were calculated from the bus data. First, the average number of pick-ups for each oriented stop was calculated (by bus route first, then a weighted average calculated to obtain a global average). This average count of pick-ups at an oriented stop was later used as the key parameter in simulating the bus routes using the Poisson distribution. Using the same calculation to assess the average number of drop-offs was initially attempted, but was decided against (a distinction from Pereira and Chwif (2018)). Thus, the second value calculated was to estimate the drop-offs: the average percent of the bus exiting at a given stop (e.g., 45% of the bus exits at stop 5 heading away from the ate). Finally, the average inter-stop time was calculated from each oriented stop to its next stop (i.e., from Stop 7 towards the gate to Stop 11 towards the gate). Of note, not all possible combinations of from one stop to another was present; in these cases, estimates based on comparable stops were used (e.g., repeating the same length of time as the previous stop, or estimating a travel time based on similar lengths on the same road).

3.2.5 Additional Taxi Data Analysis

To incorporate the taxi data into later simulation, the 93 most popular pickup locations were each assigned to the nearest directional stop, which required some assumptions made about destination. Then, dividing the demand signal by 11,808 (total number of days over which the calls were accumulated, times 24 hours in a day, times 4 average pick-ups per hour, to estimate average calls in the roughly 15 minutes between buses), this value was added to the "average" number of pick-ups at a given stop calculated from the data.

These key parameters (oriented demand signal, exit probability, and inter-stop travel time) calculated from the data allow for a simulation model to be built and assessed.

3.3 Simulation Development

All simulation for this paper was conducted in Python 3.8 using SimPy, NumPy, and Pandas packages. The general process for developing the simulation was to code a working model, then tune the model to imitate one route very well, then model the overall system behavior of all routes. Once the simulation behavior mirrored the data, then we were able to use the simulation to assess the proposed new routes.

The key component of the simulation was the use of a bus process (i.e., a function) to behave like the observed buses, including: assigning routes for a bus, inter-stop travelling time, picking up and dropping off passengers, and recording all required outputs.

3.3.1 Assumptions

There were many critical assumptions made in order to use the basic process developed for this case study.

One assumption was that a Poisson process can be used to generate a random number of passengers at each stop, based on the "demand signal" for a given oriented stop. This is not an uncommon assumption in the literature discussed previously. While we know that the number of passengers (and the rate of their arrival) is not truly independent of other stops nor of time, the Poisson random variable allows us to simulate the natural variation due to varying times of day, formations, etc.

Another assumption was that inter-stop travel time was roughly constant (from one given oriented stop to its neighbor). Experience demonstrates this is not true, but the data for travel time between stops was relatively sparse ($n \leq 30$ for any given stop), so the mean value was used to have a simple, explainable inter-stop travel time value. While normal distributions with standard deviations drawn from the data were considered, they did not generate significantly different results in the final product, and so to reduce complexity, the average travel time between stops was used.

3.3.2 Simulation Validation

The simulation was initially built so that the average number of passengers on the bus across 1,000 simulations would be within accurate within 10% of the average bus load of the Green Route. For this initial setup, only the average pick-up and dropoff probability of the Green rouet were used (as compared to the final data output: a consolidated, oriented-stop dictionary). See Figure 2 to compare average bus loads across routes and see how the base simulation estimated how full the average buses were.

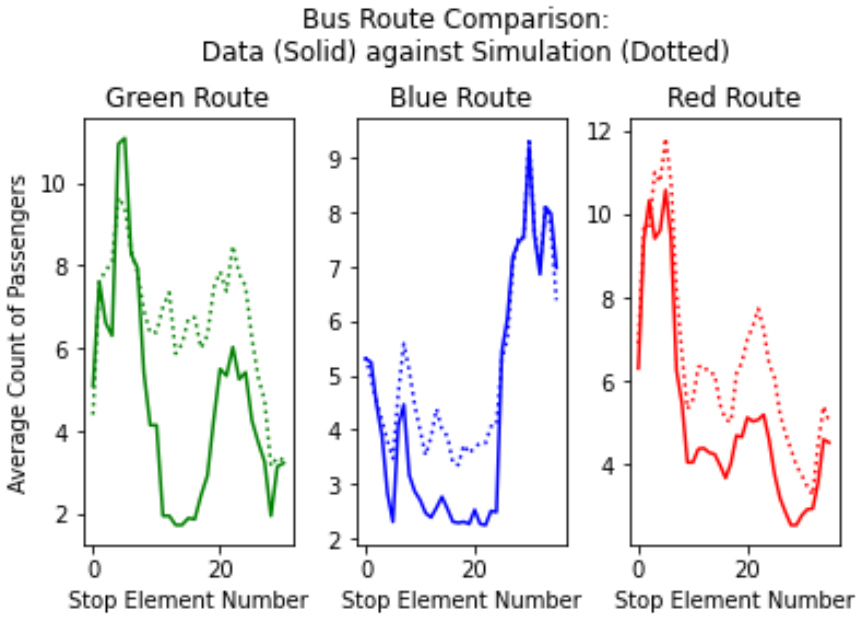

Figure 2: Comparison of average bus load along given route, solid line representing data, dotted line simulation.

With a simulation behaving comparable to one specific route, the consolidated oriented-stop data were used to generate 1,000 simulation iterations of all three bus routes to compare to the average bus load provided in the data. These models were less accurate than the Green-only model, but performed on average within 1 passenger of the data, and so were assessed to be a reasonably good fit to reality while still retaining explanatory power.

3.3.3 Metrics

In order to compare the data with simulated new bus stops, a metric had to be selected to compare different routes. While a successful bus route might be measured in practice by fewer taxi pickup calls, a higher quantity of passengers, or greater bus utilization rate, this was not meaningful in a simulation environment. For the simulation to "increase ridership", one need only visit stops more frequently, or only visit the most popular stops more rapidly, and then the simulated number of riders would increase, which would clearly not effectively model reality.

In place of this, the metric of "average rider time" was used. For each leg of the bus trip, the total quantity of passenger-minutes for that leg was calculated and added to a running total for that bus trip. At the end of the trip, the total number of passenger-minutes was divided by the total number of passengers to get the average bus trip duration. This metric could allow us to compare different route structures to explore potential increases in efficiency.

With a clear metric and functional simulation, proposed routes can be evaluated against previously existing routes to assess if they are an improvement.

3.4 New Route Proposal

Using principles inspired in the literature, including hub-and-spoke structure and minimizing overlap between routes, three new routes were developed. See Figure 3 for a visualization of the new routes.

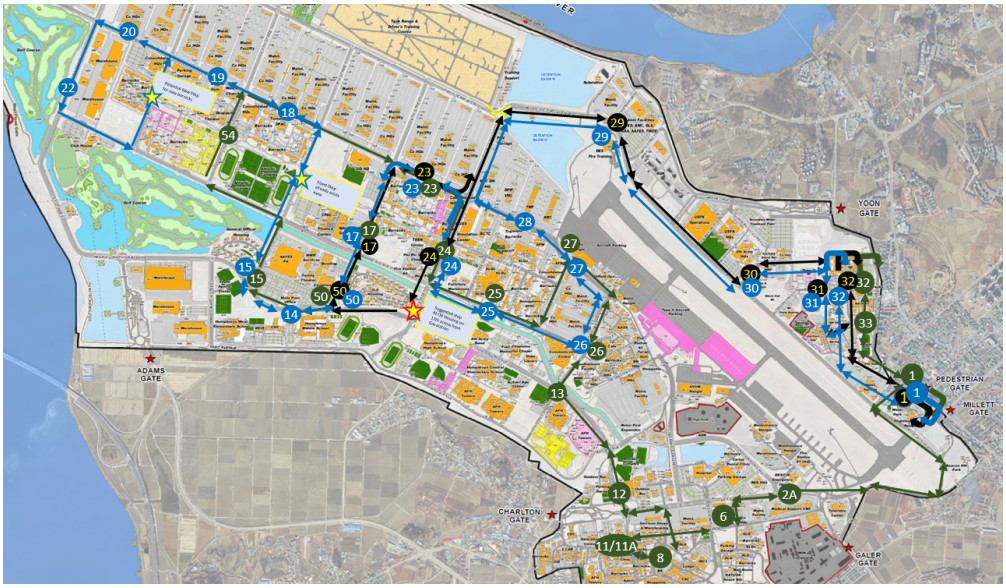

Figure 3: Depiction of new proposed routes.

To estimate average rider time for these routes, the routes were to be programmed using the current oriented stops. Where data were missing for inter-stop travel time (e.g., a new string of stops not currently serviced, or where previously used stops were skipped along a route), an estimate based on comparable average interstop travel time was used. All passenger pick-up random distributions used the average oriented stop values from data, as were percent of the bus to exit at a given oriented stop.

From these inputs to the simulation, 10,000 iterations of simulation were run, and the average bus load and average rider time were computed.

4 Results

Table 4 summarizes the major results of the simulation outputs. For each of the simulations for the original routes, 1,000 simulated bus routes were used. For the new routes, 10,000 simulations were used to estimate potential bus loads and average rider time.

	Green Route	Blue Route	Red / Black Route	Weighted Avg. Travel Time
Old Routes	12.94	12.68	12.06	12.57
New Routes	9.55	9.38	9.46	9.47

Table 1: Major average outputs from simulation.

These results, specifically the 3-minute decrease in average rider time, indicate the simulation suggests these routes to be an improvement that should offer a more competitive alternative to taxis on Camp Humphreys.

5 Discussion

The goal of this case study was to present a real-world case study of simulation, in conjunction with data collection and analysis, solving tangible problems. In this case, to reduce a large amount of Soldier and

family money spend unnecessarily on taxis on Camp Humphreys, and to use installation resources (money, personnel, and equipment) more effectively.

5.1 Data

In retrospect, the data could have provided more fidelity. The "gold standard" of collection of transportation data include passenger pick-up and drop-off pairing, which can help identify specific patterns that might be needed; but due to a variety of limitations this was infeasible. Some other issues were the sparse collections across varying times, many observations starting and ending at a non-bus depot stop, or inconsistencies in data entry. All of these aside, the data still provide an invaluable window into understanding and visualizing both the oriented demand signal of a given stop on a route and understanding what an "average bus trip" looks like for the given routes. As such, these data were the touchstone of the entire project and facilitated the project's grounding in reality.

In conjunction with this analysis, the 403rd AFSBn has also added a schedule of regular reviews to the bus route to gather data, solicit feedback, and improve their routes and bus offerings. Additionally, while the aim of the paper is not to discuss bus costs, the recommended proposals in the final routes reduced driver requirements by approximately 37%, reducing daily driver man-hours from 195 to 123.

5.2 Simulation

As discussed previously, the simulation relies on several assumptions that we know are not perfect, e.g., passengers don't truly arrive independently from one another without respect to how many are at a stop waiting. At the individual level, random passengers who arrived at a given stop will likely not depart at some other stop with a set probability; usually passengers get on a bus with an intended destination in mind. However, taken at the aggregate level, these assumptions do accurately describe the general way the body of passengers who take a bus route behave. So when taken at the aggregate level, these imperfect assumptions generally help model reality.

This case study is an example of how, from data collection through analysis and use of freely available simulation software, a small team of analysts can collect, analyze, generate options, and use simulation to help quantify and assess efficacy of those different options. It is hoped by the authors that more robust examples continue to illuminate well-defined and quantifiable options to decisionmakers.

Acknowledgements

The authors would like to thank the many individuals without whose assistance, none of this work would have been possible. Specifically, Mr. Charles Stafford (403rd AFSBn), Mr. Hyon Kun Ma (Department of Public Works), Ms. Vanessa Rowland (Army Air Force Exchange Services, AAFES), and Mr. Jeffre Nagan (United States Army Garrison-Humphreys Public Affairs) – thank you for your assistance and support in gathering the tools and data required for this study.

A Appendices

The code or data for this analysis is available from the authors upon request.

REFERENCES

López, E. C., F. Marmier, and F. Fontanili. 2019. "Bus fleet size dimensioning in an international airport using discrete event simulation". In *2019 Winter Simulation Conference (WSC)*, 464–475. IEEE.

Pemberton, S. 2020. "Optimising Melbourne's bus routes for real-life travel patterns". *Case Studies on Transport Policy* 8(3):1038–1052.

Pereira, W. I., and L. Chwif. 2018. "Generic bus route simulation model and its application to a new bus network development for caieiras city, Brazil". In *2018 Winter Simulation Conference (WSC)*, 123–134. IEEE.

Schmaranzer, D., R. Braune, and K. F. Doerner. 2016. "A discrete event simulation model of the Viennese subway system for decision support and strategic planning". In *2016 Winter Simulation Conference (WSC)*, 2406–2417. IEEE.

Shi, Q., K. Zhang, J. Weng, Y. Dong, S. Ma, and M. Zhang. 2021. "Evaluation model of bus routes optimization scheme based on multi-source bus data". *Transportation Research Interdisciplinary Perspectives* 10:100342.

Vázquez-Abad, F. J., and L. Fenn. 2016. "Mixed optimization for constrained resource allocation, an application to a local bus service". In *2016 Winter Simulation Conference (WSC)*, 871–882. IEEE.

Author Biographies

DEVON ZILLMER is an instructor in the Department of Mathematical Sciences at the United States Military Academy. His research interests include pedagogy, algebra and cryptography, Thomistic philosophy, data analysis, and modeling. His email address is devon.zillmer@westpoint.edu.

ANNA TUCKER is an instructor at the United States Military Academy. Her research interests include observability, controllability, and sustainable energy systems. Her email address is anna.tucker@westpoint.edu.

JAMES K. STARLING is an assistant professor at the United States Military Academy. His research interests include obsolescence management, optimization, simulation, and military applications. His email address is james.starling@westpoint.edu.

BRYAN HARTMAN is an operations research systems analyst stationed at Camp Humphreys, South Korea. His research interests include realization analysis, applied systems engineering, cost analysis, and military applications. His email address is bryan.d.hartman4.mil@army.mil.

PAUL KUNNAS is a computer simulations and systems integrator stationed at Camp Humphreys, South Korea. His research interests include operations efficiency, military data and cloud integration. His email address is paul.e.kunnas.mil@army.mil.