BISTRO: Berkeley Integrated System for Transportation Optimization

SIDNEY A. FEYGIN*, Uber Technologies, Inc.

JESSICA R. LAZARUS* and EDWARD H. FORSCHER*, University of California, Berkeley; Uber Technologies, Inc.

VALENTINE GOLFIER-VETTERLI*, JONATHAN W. LEE*, and ABHISHEK GUPTA*, Uber Technologies, Inc.

RASHID A. WARAICH and COLIN J.R. SHEPPARD, Lawrence Berkeley National Laboratory

ALEXANDRE M. BAYEN*, University of California, Berkeley; Uber Technologies, Inc.; Lawrence Berkeley National Laboratory

The current trend towards urbanization and adoption of flexible and innovative mobility technologies will have complex and difficult-to-predict effects on urban transportation systems. Comprehensive methodological frameworks that account for the increasingly uncertain future state of the urban mobility landscape do not yet exist. Furthermore, few approaches have enabled the massive ingestion of urban data in planning tools capable of offering the flexibility of scenario-based design.

This article introduces BISTRO, a new open source transportation planning decision support system that uses an agent-based simulation and optimization approach to anticipate and develop adaptive plans for possible technological disruptions and growth scenarios. The new framework was evaluated in the context of a machine learning competition hosted within Uber Technologies, Inc., in which over 400 engineers and data scientists participated. For the purposes of this competition, a benchmark model, based on the city of Sioux Falls, South Dakota, was adapted to the BISTRO framework. An important finding of this study was that in spite of rigorous analysis and testing done prior to the competition, the two top-scoring teams discovered an unbounded region of the search space, rendering the solutions largely uninterpretable for the purposes of decision-support. On the other hand, a follow-on study aimed to fix the objective function, served to demonstrate BISTRO's utility as a human-in-the-loop cyberphysical system: one that uses scenario-based optimization algorithms as a feedback mechanism to assist urban planners with iteratively refining objective function specification and constraints on intervention strategies such that the portfolio of transportation intervention strategy alternatives eventually chosen reflects high-level regional planning goals developed through participatory stakeholder engagement practices.

CCS Concepts: • **Human-centered computing → Collaborative interaction**; • **Computing methodologies → Multi-agent systems**; **Agent / discrete models**; *Simulation tools*; • **Applied computing → Transportation**.

*At the time of writing, Lee and Gupta were full-time employees at Uber. Feygin, Lazarus, Forscher, Golfier-Vetterli, and Bayen contributed to this work partially as Uber contractors.

Authors' addresses: Sidney A. Feygin, sfeygi@ext.uber.com, Uber Technologies, Inc. Jessica R. Lazarus, jlaz@berkeley.edu; Edward H. Forscher, forscher@berkeley.edu, University of California, Berkeley; Uber Technologies, Inc. Valentine Golfier-Vetterli, vgolfi@ext.uber.com; Jonathan W. Lee, jonny@uber.com; Abhishek Gupta, agupta@uber.com, Uber Technologies, Inc. Rashid A. Waraich, waraich.rashid@gmail.com; Colin J.R. Sheppard, colin.sheppard@lbl.gov, Lawrence Berkeley National Laboratory; Alexandre M. Bayen, University of California, Berkeley; Uber Technologies, Inc.; Lawrence Berkeley National Laboratory, bayen@berkeley.edu.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2019 Association for Computing Machinery.

Manuscript submitted to ACM

Additional Key Words and Phrases: Agent-based models, human mobility, system dynamics, big data, smart cities, digital decision support systems, urban informatics, computing with heterogeneous data, intelligent transportation systems

ACM Reference Format:
Sidney A. Feygin, 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. 2019. BISTRO: Berkeley Integrated System for Transportation Optimization. 1, 1 (July 2019), 27 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Mobility pattern modeling in the age of data science. As modern transportation systems undergo a period of intense technological evolution, researchers, practitioners, and policymakers are seeking to understand how emerging trends in digitalization, automation, electrification, and the sharing economy are, in turn, shaping the dynamics and broader impacts of human mobility in cities worldwide.

Likewise, rapid advances in computing as well as the advent of metropolitan scale data mining and pattern recognition (*i.e.*, machine learning) have, for the first time, made it possible to characterize urban traffic flows based on movement traces from millions of individual travelers. These methods fuse passively collected spatiotemporal trajectories derived from smartphone data with static census data in order to map travel patterns to sociodemographic characteristics. Specialized traffic models can thereby be trained using feature-rich representations of human mobility.

While such predictive models have proven effective in nearcasting congestion events [1], purely data-driven methods often fail to generalize when applied to the task of forecasting the effect of transportation policy strategies on future demand–particularly when anticipated changes in the urban social, physical, and technological landscapes as well as interactions between these layers cannot themselves be predicted with a high degree of certainty [2]. In other words, the co-variates of the trained model may not capture the numerous interacting physical and behavioral components (*i.e.*, *network effects*) influencing transportation supply and demand.

ABMS. Agent-based modeling and simulation (ABMS) techniques and software, on the other hand, are beginning to be used as flexible decision-support tools by planning groups, technologists, and regulatory agencies to facilitate forecasting the short- and long-term implications of transportation system interventions. That is, by simulating traveler decision-making in the context of a landscape of realistic, but counterfactual "possible worlds", stakeholders can better resolve deep uncertainty about how transportation system interventions will fare in novel scenarios. For example, ABMS enables the investigation of the extent to which citizens may adopt future and emerging transportation modes, such as e-bikes or scooters, as well as how to incentivize their sustainable use. In order to map aggregate model outputs to corresponding metrics derived from real-world observations, researchers are increasingly exploring calibration and validation protocols that make use of so-called "big data"-driven methods, such as the statistical learning techniques indicated above.

Even with these advances, selecting interventions that balance competing policy objectives for transportation remains a difficult and contentious process. Current methods to identify sets of optimal alternatives under a given scenario often involve the use of Monte Carlo methods to "try out" different input parameters. Moreover, transportation researchers and practitioners often develop scenario-specific and/or geography-specific simulation models with limited integration of the efficient and highly generalizable methods developed in the *artificial intelligence* (AI) and *machine learning* (ML) communities. Significant opportunity exists to combine the data-driven technologies emerging from Computer Science with the model-based protocols typically used in Transportation Science and Engineering.

BISTRO. The *Berkeley Integrated System for TRansportation Optimization* (BISTRO) is an open-source *Collaborative Planning Support System* (CPSS) designed to assist stakeholders in addressing the increasingly complex problems arising in transportation systems worldwide. BISTRO includes an ABMS framework and scenario development pipeline to build empirically-validated simulations of multimodal metropolitan transportation systems and algorithmically optimize system interventions that best align with policy and planning objectives. Users can deploy BISTRO to enable distributed development of algorithms that rapidly optimize a feasible set of policy and investment decisions. Once one or more desirable solutions are found, BISTRO provides a suite of analysis and visualization tools to empower citizens, transportation system planners and engineers, private entities, and governments to better understand and collaborate on developing strategies that achieve equitable access to and sustainable use of current and emerging mobility services.

The rest of this article is organized as follows: Section 2 provides a brief overview of the current state of transportation planning, ABMS tools, and transportation optimization tools, respectively; Section 3 presents BISTRO, covering the system architecture, scoring function design, inputs, outputs, analysis capabilities, and performance characteristics; Section 4 details the initial pilot study, updates to BISTRO based upon the pilot, and a few algorithmic solution approaches; and Section 5 offers a short conclusion.

2 BACKGROUND

2.1 Transportation Planning and Policy Decision Support Systems

Transportation planning in the era of machine learning. In the United States, the predominant form of data-driven planning to investigate the effects of policy and infrastructure changes on transportation demand and externalities is done via Metropolitan *Regional Transportation Plans* (RTPs) and State *Long-Range Transportation Plans* (LRTPs). These plans are forward-looking, long-term (20+ year time horizons) and have been a federally mandated task since the Federal-Aid Highway Act of 1962. Recently, states and *Metropolitan Planning Organizations* (MPOs)—the entities tasked with producing LRTPs every four-to-five years—have been shifting towards vision- and/or goal-driven plans evaluated with performance assessments [3]. Transportation and city planners and stakeholders who work in and around local, regional, state, and federal policy-making are likely quite familiar with LRTPs and associated processes, such as *Transportation Improvement Programs* (TIPs), *Regional Housing Needs Assessments* (RHNAs), and others.

During the RTP process, localities will inform regional entities of their housing needs, production, projected traffic and demographic shifts, among other details. Cities will usually host extensive public engagement processes to hear from citizens about perceived needs and concerns. Regional planners will then combine all of this information into a regional-scale land use and transportation model, analyzing the collective projected outcomes of no-build and a variety of alternative investment and programming strategies on the region's population.[1] This type of alternatives analysis then informs a draft plan, which acts as a guiding tool for the region to allocate funding. During the entire process, MPOs will usually involve the public whenever possible, seeking feedback both informally (prior to the release of the draft plan) and formally (during the public comment period once the draft plan is released). Once the comment period is closed and the MPO feels it has adequately taken into account citizen concerns, it will release a final plan and move forward with environmental approvals processes, as applicable.

[1] In the past the modeling process often consisted of four main steps: 1) trip generation to and from all analysis zones, 2) trip distribution (or matching origins and destinations, often using a gravity model), 3) assigning traveler mode choice based upon individual preferences and alternative characteristics, and 4) route assignment, of trips onto physical network links; this is referred to as the four-step model. Many MPOs and other agencies are moving towards activity-based, or person-centric models of daily activity, rather than zone-based approaches [3]

Although many MPOs conduct extensive public outreach processes to inform their plans[4], the nature of the models underlying the forecasting process can still be frustratingly opaque to the general public and explanations of the inner workings of the modeling process are not presented during public collaboration meetings[5]; questions regarding the validity of models has led to lawsuits over the lack of publicly available information regarding them[2][6–8].

Digital collaborative transportation planning. Recent advances in modeling from both the public and private sectors have begun to promote publicly available and openly collaborative simulation platforms [3, 9–11]; however many of these tools have been developed for narrowly defined purposes or bounded within specific geographies, limiting their generalizability. As discussed above, many MPOs throughout the United States maintain highly detailed model libraries for their regions–capable of analyzing short and long-term urban development–in support of RTP development. While these models are used as the basis of regional planning and community engagement, the underlying data and software libraries are often not readily available in open source repositories. This access impedance limits the seamless transfer of policies and ideas across and even within agencies, often resulting in one-off models that are expensive to develop and restricted to one specific region of study[12, 13].

2.2 ABMS of Transportation Systems

Understanding how urban systems operate and evolve has been a major focus of transportation engineers, urban planners and geographers. Trip-based methods, such as the traditional four-step model used by MPOs, are not represented using human decision-making, which impairs their ability to forecast how incentives and policies impact behavior changes. Some MPOs have begun to adopt more a more sophisticated activity-based approach [14]. Activity-based models represent more comprehensive links between activity scheduling, mode choice, social interaction, and spatiotemporal constraints [14]. Agent-based models and simulations (ABMS) of transportation systems have been demonstrated as capable of subsuming many of these objectives. Consequently, urban planners have increasingly adopted agent-based tools in order to forecast long-term effects of transportation system changes due to anticipated derangements of existing travel patterns expected to be caused by the introduction and adoption of future and emerging technologies [2]. This section presents background on ABMS of urban transportation systems with a focus on concepts and frameworks relevant to BISTRO.

Background on ABMS. ABMS has long been used in economics, sociology, and the biological sciences as a method to study so-called *emergent properties* or *emergent behaviors* in complex natural and social systems [15, 16]. These are distinct and surprising macro-level phenomena that arise when individual decision-making units, or, *agents*, interact with each other according to a simple set of rules. In models of physical systems, these interactions are often mediated via a network, situating the behavior of agents relative to one another in space and time. Agents typically possess imperfect and limited information about the overall state of the system and history of its evolution [16]. By imbuing agents with preferences and the goal of maximizing personal utility, researchers can observe how choices that appear optimal to individual agents in resource-constrained physical environments can result in negative externalities that, in aggregate, lead to suboptimal equilibria.

When applied to research concerning the interactions between human and physical geographies, ABMS approaches can be viewed as virtual laboratories within which investigators may conduct experiments that would be either ethically or practically impossible to perform on real-world populations [16]. Many ABMS frameworks are intended to generalize

[2]Many such suits often center on challenges to environmental impact statements/reports, as the environmental process is a common avenue for public engagement (positive or adversarial) regarding transportation plans or projects in the United States.

across a variety of locations. Adapting models to new locations is accomplished by *calibrating* simulated behavior to ground-truth data representing a snapshot of the state of the world. Calibration often involves tuning a set of parameters that influence individual and system-level properties until aggregate model outputs match the corresponding statistics derived from empirical observations.

Once calibrated, ABMS allows for the evaluation of counterfactual *scenarios* [2, 16]. A scenario is a simulation that implements a unique set of circumstances that differs in some way from a *base case*. Examples of scenarios used by agent-based models of urban geographies and, in particular, the transportation system include alteration of model parameters representing population or employment growth, alteration of the transportation network such as expected or unexpected road network restrictions due to sporting events, inclement weather, or traffic collisions, as well as the introduction of new modes of transportation such as autonomous vehicles or increasing the supply of existing shared modes such as e-bikes. A simulated evaluation of each scenario can reveal the result of these as well as many other counterfactual realities on the decisions of individual agents, and thus, the aggregate impacts of possible changes in the choice environment.

In the past three decades, several multi-agent frameworks such as TRANSIMS [17], MATSim [18], SUMO [19], and POLARIS [9] have been developed and widely adapted for numerous applications in transportation and land use planning and research [2, 20, 21]. MPOs often couple urban development simulation models[3] with microscopic agent-based transport models to better understand how predicted changes in population growth, land-use, real-estate development, and resource markets will co-evolve with changes in the transportation system [24, 25]. ABMS of transportation systems can take different population configuration files as inputs, giving planners and modelers the ability to simulate how long-term urbanization processes can be shaped by the daily transportation decisions of individuals (possibly in the presence of alternative policy interventions and new mobility technologies) [25, 26]. Increasingly, these open-source platforms are enabling use of publicly available data to create transparent and replicable input preparation pipelines, reducing the cost and effort of a variety of urban planning tasks [27].

MATSim. MATSim is an ABMS framework developed by teams at ETH Zürich and TU Berlin [18]. MATSim enables modeling of the travel behavior of millions of individual agents, representing a synthetic population of urban travelers. At the heart of MATSim is a co-evolutionary algorithm that iteratively executes, evaluates, and mutates (i.e., replans) the activity schedules of agents (see Figure 1), creating realistic congestion patterns as agents compete for limited space on a virtual road network.

The agent population is initially characterized by a set of unrealized plans (one for each agent) consisting of the start times, types (e.g., “Home”, “Work”, “Shopping”, “School”, etc.) and locations of various significant activities. A mobility simulation (`MobSim`) executes these plans on a virtual road network. While, at first, agents only drive and/or walk to activities, additional modes may be introduced during plan mutation (explained below). Each agent’s plan is then scored according to a generalized daily utility function.[4] Copies of evaluated plans are

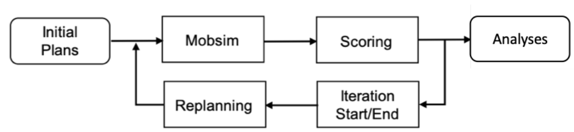

Fig. 1. Conceptual process of MATSim. It iteratively evaluates and mutates a proportion of agent plans until the utility of plans no longer improves. At this point, the system is said to have reached a *stochastic user equilibrium*. For further details, see [18].

[3]Examples of these include the *Integrated Transportation, Land Use, and Environment* model (ILUTE, [22]) and UrbanSim [23].
[4]Utility in MATSim refers to the utility experienced when executing an activity-travel plan. It is measured by a linear model that assigns negative value to time spent traveling and positive value to time spent at activities.

stored in a limited-size array, representing the agent's memory. At the start of the subsequent iteration, a portion of agents in the population are chosen to have a randomly selected mutation strategy applied to a plan drawn from each of their memories. Examples of plan mutations include changing activity start time, mode of travel for a tour (or subtour), and selecting a different route based on previously experienced link travel times.

The algorithm converges once agents are no longer able to improve the utility of plans in their memory, at which point MATSim produces a series of statistics and outputs describing the aggregate performance of system components as well as a snapshot of all events that occurred over the course of the simulation. Events can be processed to derive the actual paths and travel times realized by each agent and each vehicle, as well as a host of other data.

BEAM. The *Behavior Energy Autonomy and Mobility* (BEAM) framework is a multi-agent travel demand simulation framework developed at *Lawrence Berkeley National Laboratory* (LBNL) [28]. While the overall learning and traffic assignment mechanism is similar to MATSim's co-evolutionary algorithm, added functionality in BEAM is specifically focused on helping users understand the impacts of new and emerging travel modes on limited capacity resource markets. This subsection describes essential features of BEAM that led to its selection as the core simulation engine of BISTRO.

Due to the expanding variety of mobility alternatives, multiple factors driving down the propensity to drive alone (such as environmental awareness, gas prices, and availability of novel and more convenient or less expensive alternatives), as well as real-time information provided via smartphone apps, travelers are increasingly combining multiple travel modes into one trip. BEAM is closely integrated with the transit service capabilities of the R5 routing engine, which include *General Transit Feed Specification* (GTFS) file processing and routing based on multiobjective variations of the RAPTOR algorithm [29]. Transit may be combined with other modes modeled in BEAM such as autonomous vehicles, on-demand rides, e-bikes, and scooters, enabling agents to make realistic, multimodal mobility decisions.

In order to provide agents with information about the time and monetary costs of different travel options, R5 computes the lowest generalized cost path (based on travel time estimates from the mobility simulation) for the corresponding mode(s) available to the agent for the trip.[5] The probability of selecting a route returned by R5 is represented in BEAM according to a multinomial logit model [31, 32]. That is, among several distinct alternatives, agents are exponentially more likely to select the alternative that maximizes their enjoyment of important activities while reducing time and money spent traveling between activity locations. BEAM ensures realistic variation in agent preferences by assigning sociodemographic attributes to agents in accordance with statistical distributions derived from census data during population synthesis.

Unlike the replanning mechanism in MATSim, agents in BEAM can adapt to changing conditions during an iteration according to what is known as a *within-day* or *online* model. Thus agents can make unplanned and time-sensitive choices about how to maximize the utility of their travel plans while competing for limited resources that vary in availability over time. For example, an agent that chooses a transit mode may be denied access to an overfull bus, requiring the agent to make a mid-trip change to their itinerary. The BEAM software architecture addresses the performance and complexity challenges of integrating new models of within-day dynamics by implementing agents as *actors*, as defined within the *actor-based model of concurrency*.[6]

[5]For detailed information about the R5 router, see [30].

[6]Like objects in the object-oriented programming paradigm, actors encapsulate state and behavior. However, unlike the object model, actors do not share computer memory. Instead, each actor encapsulates its own thread of execution and interacts with one other actors using messages. An actor may send message to other actors without blocking. Each actor processes messages synchronously in the order received; however, computation is scheduled asynchronously over multiple actors. Thus, the actor-based model of computation obviates the need for locking mechanisms commonly used

2.3 Simulation-based Optimization of Transportation Systems

2.3.1 Optimization-based formulation of the planning problem. The problem class solved by the BISTRO framework can be characterized as simulation-based optimization of large urban transportation systems. It can be symbolically formulated as an optimization problem:

$$\underset{\vec{d}\in\mathcal{D}}{\text{minimize}}\ f(\vec{d},\vec{x};\vec{z}) \equiv \mathbb{E}[F(\vec{d},\vec{x};\vec{z})] \tag{1}$$

constrained by simulation outcomes and design constraints, *i.e.*,

$$\begin{cases} \vec{x} = B(\vec{d};\vec{z}), \\ g(\vec{d};\vec{z}) = 0, \end{cases} \tag{2}$$

respectively, where the objective, f, is defined as the expected value of a stochastic performance measurement function, F. The deterministic decision vector, $\vec{d}$, is chosen from a search space $\mathcal{D}$, which may be continuous, categorical, combinatorial, or conditional. In the BISTRO context, the decision variables, $\vec{d}$, are the user-defined inputs that control policy levers within the transportation system. The exogenous variables, $\vec{z}$, are the configuration inputs that determine the parameters of the population synthesis, the parameters of the transportation network, and the parameters governing supply of transportation services. The endogenous variables, $\vec{x}$, are the outcomes of the simulation run using $\vec{d}$ and $\vec{z}$ as input. The vector $\vec{x}$ contains the details of agent and vehicle movements throughout the simulation run, such as mode choices, travel times, travel costs, and vehicle path traversals, that were realized during the simulation run, *i.e.*, $\vec{x} = B(\vec{d};\vec{z})$, where B represents the BEAM simulator. It is assumed that the iterative simulation process described in Section 2.2 has achieved stationarity.[7] In BISTRO, the constraint function g is defined according to business rules that ensure interpretability and realism in the optimal solution. Finally, F, is computed as a convex combination of the score components that guide solutions towards the *system objective* (as defined in Section 2.1).[8]

2.3.2 Optimizing complex simulated systems: challenges and approaches. When adapting transportation policy to counterfactual or hypothetical future scenarios, retaining parameters responsive to proposed incentives and other behavioral and infrastructure interventions is highly desirable. As described in Section 2.2, agent-based micro- or meso-scale simulations of transportation systems model the interdependent choices of rational individuals as they navigate virtual representations of physical and human geographies. Calibrating such models to high-resolution GPS traces and other sensor data embedded in infrastructure makes them highly suitable for evaluating the outcomes of location-specific policy alternatives. The trade-offs in accounting for the heterogeneous preferences of millions of agents, are that 1) the simulator is expensive to evaluate for different settings of $\vec{d}$, and 2) the complex relationship between network dynamics and agent behavior lead to stochastic, non-convex specifications of performance measure, F. Consequently, the efficient gradient-based methods used to optimize closed-form relaxations of mobility dynamics as well as data-driven models derived from historical movement patterns do not apply [34]. Instead, generalized *stochastic*

to synchronize state among interdependent objects. Consequently, reasoning about agent behavior using actors can allow researcher developers to focus on implementing novel models and applications rather than debugging threads and locks [33].

[7]Individual optimization algorithms may relax this constraint in order to reduce compute time while potentially trading off reduced accuracy or increased stochasticity of simulation output statistics.

[8]Due to variable amounts of nondeterminism and stochasticity inherent in ABMS, given fixed $\vec{d}$ and $\vec{z}$, the distribution of f can be approximated using n realizations of F as $\hat{f}\left(\vec{d},\vec{x};\vec{z}\right) = \frac{1}{n}\sum_{i=1}^{n} F_i\left(\vec{d}_i,\vec{x}_i;\vec{z}\right)$. In practice, optimization usually proceeds with $n = 1$ in order to identify promising (i.e., close to optimal) subsets of $\mathcal{D}$; however, when reporting final scores, one must carefully select n such that variability in output values is adequately captured.

optimization (SO) algorithms treat the simulator as a black box. Commonly used derivative-free SO approaches include grid search, random search, ranking and selection, metaheuristic, and *metamodeling* techniques [35, 36].

Metamodeling algorithms encompass a broad class of simulation-based optimization approaches. These approximate F using a *surrogate model*, $\mathcal{Q}$ that is less costly to evaluate. Flexible and computationally tractable representations such as polynomial splines are able to approximate any objective function; however, many simulation runs are still required to accurately fit the response surface of the underlying system [35, 36].

Sequential model-based optimization (SMBO) is a general metamodeling formalism that, given a history of previously evaluations, $\mathcal{H} = \left\{\left(\vec{d}_1, y_1\right), \ldots, \left(\vec{d}_i, y_i\right)\right\}$, of observations $y_i = F\left(\vec{d}_i, \vec{x}_i; \vec{z}\right)$ at sample points in $\mathcal{D}$, selects the optimal next point $\vec{d}_{i+1}$ based on an approximation of F. To initialize SMBO, a small set of samples, $\left\{\vec{d}_1, \ldots, \vec{d}_i\right\}$ from $\mathcal{D}$ are selected using various experimental design techniques (e.g., random or Latin hypercube sampling). For each $\vec{d}_i$, evaluations of the expensive objective function, F form an observation, which, together with $\vec{d}_i$ are appended to a historical dataset $\mathcal{H}$. Once $\mathcal{H}$ is initialized, SMBO then proceeds iteratively: First, a regression model, $\mathcal{Q}$, is fitted to the current dataset, $\mathcal{H}$, yielding a surrogate model for F at the current iteration, which may be denoted $\mathcal{Q}_i$. Based on Q_i, the next input, $\vec{d}_{i+1}$ to F is selected by optimizing an *acquisition function*, $\alpha : \mathcal{D} \mapsto \mathbb{R}$ over $\mathcal{D}$, which measures the utility gained from evaluating F at $\vec{d}_{i+1}$. Following evaluation of $F\left(\vec{d}_{i+1}, \vec{x}_i; \vec{z}\right)$, $\mathcal{H}$, is updated as $\mathcal{H} = \mathcal{H} \cup \left(\vec{d}_i, y_i\right)$. The SMBO process continues until a predetermined time or computation budget is exhausted.

SMBO techniques are typically distinguished by the forms of the surrogate model, Q, and the acquisition function, α. In Bayesian optimization (BayesOpt), a Gaussian Process (GP, [37]) is typically used to model a prior over Q, which, at each iteration, is updated using previously observed data $\mathcal{H}$ to give a posterior predictive distribution $p\left(y \mid \vec{d}, \mathcal{H}\right)$ [38, 39]. Several methods using GPs as surrogate models may be distinguished according to the form of the covariance kernel parameterizing the GP [38, 40]. In lieu of GPs, BayesOpt algorithms have also used random forests [41] and *tree-based Parzen estimators* (TPE) [42, 43] as priors over Q. Acquisition functions are chosen to balance *exploration* and *exploitation* in the sample domain. The most common acquisition function used by these methods is based on an expected improvement criterion [44]; however, newer methods use variations on knowledge gradients [45, 46].

Many SMBO algorithms can run trials in parallel, which may yield reduced wall clock time (although the total number of trials required may be identical to that used by sequential implementations) [47]. One method to further reduce the running time of SMBO *trials* when model evaluations require an inner iterative loop to achieve stationarity (as in BISTRO) is to incorporate an early stopping rule for simulation evaluations that are likely to eventually be extremely suboptimal. These approaches comprise "freeze-thaw Bayesian Optimization" [48].

Recent efforts in transportation science and operations research have also sought to develop tractable simplifications of scenario-based optimization of simulated urban transportation systems. One vein of research concentrates on deriving deterministic analytic equations describing system dynamics at equilibrium from static information, $\vec{z}$ (*e.g.*, network topology and bus schedules) to inform purely functional metamodels [49–51]. For example, [34] combine a computationally tractable model of congested traffic based on queuing theory with a detailed local approximation using a linear combination of basis functions from a parametric family. Alternatively–and analogously to the "freeze-thaw" Bayesian optimization setting highlighted above–some approaches use information about the *process* by which the stochastic simulation achieves stationarity to develop techniques that rapidly evaluate different settings of the decision variable vector, $\vec{d}$, while avoiding the need to reach convergence [52, 53].

The present work refrains from prescribing a single best approach to solve optimization program Equations (1) and (2). Instead, the intent of BISTRO is to enable *replicable* future research in this area by providing a platform and problem setting that is generalizable across different planning contexts as well as approachable and of research interest to the ML/AI community. Problem characteristics such as the propensity for competing metrics to be present in system objectives, pre-emptive stopping of inner optimization loops, the high dimensionality of the search space, and the potential for hybridization of functional and physical metamodels are expected to provide challenging, scalable, and, critically, explainable solution approaches. Algorithms combining data-driven dimensionality reduction techniques as well as efficient experiment design can result in repeatable protocols to effectively constrain more general local and global search techniques.

3 BERKELEY INTEGRATED SYSTEM FOR TRANSPORTATION OPTIMIZATION (BISTRO)

BISTRO is a new analysis and evaluation platform that works in concert with an ABMS (BEAM) to enable the open-sourced development and evaluation of transportation optimization methods in response to given policy priorities. This section gives an in-depth description of the BISTRO framework and all of its major components, providing an overview of their purpose, use, and functionality as well as calling attention to the most novel aspects of its design.

3.1 System Architecture

As indicated in Section 2.3, BISTRO implements elements of the travel demand planning process coupled with components of an automated simulation-based optimization system. This section describes the high-level overall architecture of BISTRO (depicted in Figure 2), focusing on the conceptual distinction between features relevant to scenario developers and those more appropriate to algorithm designers.

A BISTRO run environment is configured using a set of fixed input data defining the required transportation system supply elements (*e.g.*, road network, transit schedule, on-demand ride fleet) and demand elements (*e.g.*, synthetic population, activity plans, and mode choice function parameters). Precisely which aspects of the virtual transportation system should be represented in the simulation model depends on the strategic goals and system objectives defined as part of the planning and analysis process motivating a particular BISTRO use case.

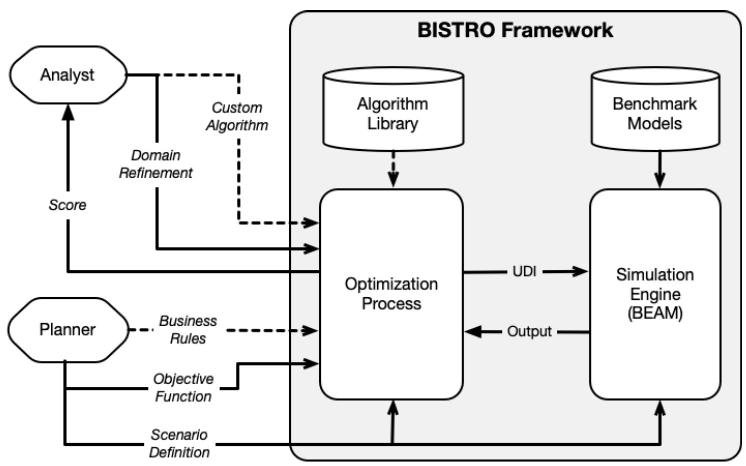

Fig. 2. BISTRO software architecture, illustrating how the optimization process modulates the flow of information between the BEAM simulation as well the two primary user types. The distinction between the *planner* and the *analyst* is critical in that we do not expect the analyst (an expert in applied ML/AI-based optimization methods) to have transportation or planning background, yet still they should be able to develop generalizable algorithms that can be used to optimize transportation system objectives set by the planning organization.

A boundary separates external, exogenously defined inputs from the BISTRO simulation optimization pipeline. Outside of the boundary, the *user-defined inputs* (UDIs) represent the investment, incentive, and policy levers applicable to and available for the study at hand. Concretely, algorithm developers encode solutions as numeric values that represent vector-valued variables controlling aspects of the initialization and evolution of the simulation. For example, a UDI that alters frequency of buses on a route must specify a target transit agency, a route, a start time, an end time, and the desired headway.

While BISTRO maintains a library of available interventions compatible with BEAM, scenario designers, policy makers, and other stakeholders will often want assurance that infeasible, regressive, or otherwise undesirable input combinations are prevented from being selected as "optimal." Together with syntactic and schematic validation of inputs, flexibly-defined *business rules* can effectively act as constraints on the search space—enhancing the interpretability and, thereby, the rhetorical and communicative value of BISTRO-derived solutions.

Just as UDIs from previously conducted BISTRO-based studies are actively maintained and made available to scenario designers, the BISTRO community contributes to a growing library of recommended optimization algorithms that, when evaluated across multiple BISTRO benchmark scenarios, demonstrate desirable performance characteristics. Thus, users lacking resources or expertise to develop optimization routines in-house can still benefit from what, we anticipate, will be cutting-edge research on algorithms and strategies to optimize the simulation of demand-responsive cyberphysical infrastructure.

Project owners of BISTRO deployments may work with stakeholders to develop representative models that will be used to benchmark optimization algorithms. Enabling a well-defined benchmark mechanism permits data on the performance of user-supplied algorithms to be compared. These comparisons can be used to assist in identification of design patterns and computational strategies that advance the state of the art in simulation-based optimization of urban transportation systems.

3.2 Scoring Function Design

Transportation system intervention alternatives are scored in BISTRO based on a function of score components evaluated using *key performance indicators* (KPIs) of the simulation. KPIs emulate common operational, environmental, and social goals considered by transportation planners and policymakers when evaluating the broader impacts of transportation policy and investment. BISTRO project planners may select KPIs to include in the scoring function from an existing library of options, or may choose to develop additional KPIs, as appropriate, for the goals and system objectives of the project. Additionally, the form of the scoring function may be designed by the analyst in consultation with the project planner.

3.2.1 Key Performance Indicators.

KPI overview. There are two general types of KPIs developed in BISTRO: 1) KPIs that measure the operational efficiency of the transportation system (*e.g.*, *vehicle miles traveled* [VMT], vehicle delay, operational costs, revenues) and 2) KPIs that evaluate the experience of transportation system users (*e.g.*, generalized travel expenditure, bus crowding experienced, accessibility). KPIs may be aggregated or disaggregated into score components to support particular policy objectives. For example, the accessibility KPI (detailed below) may be disaggregated by activity type, time period, mode used, and/or sociodemographics in order to evaluate the distributional equity of access provided across different opportunities at varying times of day and/or across population segments of concern.

In practice, any KPI that may be evaluated from the set of output variables (see Section 3.4) produced by a BISTRO simulation run may be included as a score component in the scoring function. However, careful consideration of candidate KPIs must include an evaluation of the sensitivity of the metric to the UDIs of interest as well as the efficiency of the KPI in providing the desired feedback regarding the optimality of outcomes of alternative UDI values. For example, *person miles traveled* (PMT) is a commonly used metric in transportation system performance measurement to gauge the amount of mobility delivered by the system. Yet, PMT is highly invariant within a scenario in BISTRO due to the fact that agent plans are fixed, thus agents will make the same trips regardless of the UDI values and the miles traveled by each agent will only vary in so much as the networks available for each mode offer more or less direct paths to travel from the origin to destination of each trip.[9]

Modal split, the distribution of transportation modes used across a set of trips, is another commonly used metric that often serves as an indicator of the sustainability of the distribution of demand across available modes in a transportation system. Reduction of the modal split of single occupant vehicle use, for example, is often used as a goal for meeting an objective related to reducing congestion and/or GHG emissions relative to the quantity of trips made. However, BISTRO enables the measurement of more precise goals that directly measure the system objectives. Rather than including the modal split of single occupant vehicle trips in the scoring function, one could use a direct measure of vehicle occupancy, delay experienced, and/or GHG emissions produced in the scoring function. Nevertheless, the BISTRO visualization suite (see Section 3.4) enables the user to analyze the relationship of the optimized UDI outcomes based on a particular scoring function with metrics such as modal split that provide additional intuition of the aggregate system impacts.

Implemented KPIs. The following items represent categories of KPIs that have been developed and implemented in BISTRO at the time of publication of this article:

(1) *Accessibility.* While the term accessibility takes on a variety of meanings in different contexts, in an urban transportation planning setting, accessibility has often been defined as a measure of the ease and feasibility with which opportunities or points of interest can be reached via available modes of travel. Although there are many ways to measure accessibility, it is quantified in BISTRO as the average number of points of interest (of a specific type of activity) reachable within a given duration of time. Functionality is also provided to measure mode-specific accessibility as the sum of the average number of points of interest reachable from network nodes by car or using public transit, within a specified amount of time during specific time periods.

(2) *Generalized Transportation Cost Burden*. The socio-demographic and spatial heterogeneity of travel behavior within BISTRO enables a variety of equity-focused impact analyses. One such metric that is applicable in studies such as the benchmark discussed in Section 4.1 for which there is limited intuition about both the composition of the population demographics and the spatial distribution of resources (*e.g.*, transit access, car ownership, etc.), is the average generalized transportation cost burden, based upon the income of each household. Generalized transportation cost for a particular trip is computed as the sum of the travel expenditures of the trip (costs of fuel and fares minus incentives, as applicable) and the monetary value of the duration of the trip; the monetary value of the trip duration is calculated by multiplying total duration by the population average *value of time* (VOT).[10] Generalized transportation cost is converted into generalized transportation cost burden by dividing the generalized cost by the household income of the agent completing the trip. The average generalized transportation cost burden is thus computed for all work trips and all secondary trips separately. Although this is an aggregate measure and does not examine the changes in outcomes for specific population groups, the means to pay for each household is taken into account. Furthermore, as incomes vary widely, and because of the fractional sum,

[9]For example, a transit mode choice for a particular trip may result in more PMT than a walk mode choice for the same trip, as the sidewalk network may enable a more direct path.

[10]The population average VOT is used so as to avoid inequities that may arise from valuing the time of higher income agents higher than that of lower income agents. Using the average VOT ensures that the time of all agents is valued equally.

minimizing the expenditure of lower-income households will result in larger score improvements than for higher-income households, which sends the directionally correct signal to contestants.

(3) *Bus Crowding*. The *level of service* (LoS) experienced by public transit passengers has a direct influence on short- and long-term demand for public transit service. In the short-term, passenger demand for a particular transit line is dependent on the time and cost of alternative travel options. Thus, the frequency and service period of transit service determines the availability, wait times, transfer times, and in-vehicle times of a prospective transit trip and thus the utility of that trip in comparison to the same trip completed with alternative transportation modes. In addition, the available capacity on a transit vehicle affects whether or not the passenger can board transit at the desired time. Furthermore, the comfort afforded by the available space on a transit vehicle has long-term effects on transit demand, as passengers internalize their experience during many transit trips over time and develop an additional aversion or affinity to transit based on their expectation of the LoS. Upon experiencing the discomfort of an overcrowded transit vehicle for the same 'trip' (*e.g.*, a traveler's 8 am home–work morning commute trip), a traveler will come to expect that LoS when considering whether to take transit for that trip in the future. Though the LoS may be measured in BISTRO by any one of the factors mentioned, BISTRO includes a ready-made example of a LoS KPI related to passenger comfort: average bus crowding experienced. This metric is computed as the average over all transit legs of the total passenger-hours weighted by VOT multipliers corresponding to the load factor (the ratio of total passengers to the seating capacity) of the bus during the leg.

(4) *Vehicle Miles Traveled (VMT) and Delay.* The BISTRO KPI library includes three examples of congestion score components that provide insight into the destination- or opportunity-independent level of mobility on a network, the overall network performance, and efficiency: total VMT by all motorized vehicles in the transportation system, total vehicle delay, and average vehicle delay experienced per passenger trip. Total vehicle delay is calculated as the sum over all path traversals of the difference between the realized duration and the free flow travel time of the traversal. Vehicle delay experienced per passenger trip is calculated as the total difference between the realized duration and free flow travel time of all legs of a trip completed by modes subject to congestion.

(5) *Financial sustainability.* Most system interventions will have some impact on the flow of funds in or out of the transportation system. Thus, the inclusion of an interpretable KPI that helps stakeholders understand the general financial impacts of such interventions is necessary. The financial sustainability metric provided in the BISTRO KPI library is the sum of all public transit fares collected minus all incentives distributed (if any) and all operational costs of the public transit system. The operational costs include the total costs of fuel consumed, and hourly variable costs of bus operations (see Table 2 for an example of operational costs). Hourly variable costs include estimated labor, maintenance and operational costs. The rates for each of these factors is specified in the vehicle fleet configuration variables. In the event that a BISTRO project does not alter public transit service, the operational costs may be omitted from the KPI, if desired. The purpose of the financial sustainability KPI is to shed insight on the tradeoffs from varying transit service LoS while varying incentives for transit and other modes.

(6) *Environmental sustainability.* The environmental sustainability of a transportation system intervention may be measured as the local and/or global impacts of the system. While emissions may be estimated using VMT and vehicle fuel efficiency, BEAM enables estimation of emissions directly from the fuel consumption estimates based on the realized speeds traveled by each vehicle throughout a simulation run.[11] The VMT-based fine

[11]For more information on the methodology followed to estimate fuel consumption, please refer to the BEAM documentation https://beam.readthedocs.io/en/latest/index.html.

particulate emissions ($PM_{2.5}$) KPI captures local environmental sustainability via total $PM_{2.5}$ emissions produced by all motorized vehicles during the simulation. Using criteria pollutants, specifically particulate matter running exhaust emissions factors, presents a mileage-based measure of local air quality impacts based upon vehicle type. Additional local emissions KPIs may easily be included using the appropriate emissions factors.[12]

In the updated ("New KPIs") objective function, details of which can be found in Section 4.2, a *greenhouse gas* (GHG) emissions KPI allows the optimization to explicitly account for fuel-consumption-based global environmental sustainability. It is important to note that the GHG emissions KPI will be correlated with VMT and fine particulate emissions. Thus, inclusion of all three KPIs creates a suite of environmental sustainability metrics that may apply disproportionate weight on environmentally-related objectives, which may or may not be desirable for certain policy agendas. BISTRO project owners may choose to apply scaling factors to balance the influence of the environmental sustainability score components, as described in the following subsection.

3.2.2 Scoring Function. The BISTRO scoring function serves as the objective function by which the UDIs are optimized. The selection and/or definition of the objective function is left to the decision of the project owner, according to project directives; herein, a general structure is defined to facilitate users in the creation of custom objective functions. Multiple project objectives (referred here specifically as *score components*) may be included in the scoring function, either as individual elements within a vector of scalar-valued score components to be minimized or as parameters to a function that aggregates the objectives into a one-dimensional scalar score. The score components are computed as the normalized ratio of the value of the corresponding KPI in the given simulation run to the value of the same KPI in the *business-as-usual* (BAU) run.[13] The improvement ratios are normalized using KPI values produced by a randomized sample of the UDI space, the size of which can be defined by the BISTRO project owner. This normalization (depicted graphically in Figure 3) accounts for differences in variance across KPIs, thus allowing the score components to provide meaningful feedback on the improvement achieved for each KPI relative to the distribution of the ratios of KPI to BAU produced by the random search. The composite score is thus a function of the normalized relative improvements of the candidate input to the BAU in each metric, as follows:

$$F\left(\vec{C}_s, \vec{K}, \vec{\sigma}, \vec{\mu}, \vec{\alpha}\right) = f\left(\vec{z}, \vec{\alpha}\right), \tag{3}$$

where $\vec{K}$ is the vector of all KPIs evaluated for a given set of inputs, $\vec{C}_s$; $\vec{\mu}$ and $\vec{\sigma}$ are the vectors of normalization parameters; and $\vec{z}$ is a vector of each KPI's z-scores, *i.e.*,

$$z_i = \frac{\frac{K_i(C_s)}{K_i(C_{BAU})} - \mu_i}{\sigma_i}, \tag{4}$$

for the i-th KPI. The value of the i-th score component in the BAU case is simply $K_i(C_{BAU})$.

The default objective is to minimize the composite score function, since an increase in many of the score components actually represents a scenario that is worse than the *status quo* (*e.g.*, decreasing VMT over BAU results in a lower unscaled score than increasing VMT). To maintain consistency in this regard, the scoring function may include an additional parameter $\vec{\alpha}$ to allow for transformation of score components that are positively related to desirable outcomes (*e.g.*, improvements in accessibility). For example, if the scoring function takes the form of a sum over all score components, the parameter $\vec{\alpha}$ may be used as a coefficient of each score component that determines whether the component will be

[12]For more information on the methodology followed to develop this metric, please refer to the California Air Resources Board documentation, https://www.arb.ca.gov/cc/capandtrade/auctionproceeds/cci_emissionfactordatabase_documentation.pdf.

[13]In the BAU of a given scenario, the simulation is run without alteration from the initial configuration of that scenario.

summed or subtracted, as follows:

$$\alpha_i = \begin{cases} -1 \text{ if it is desirable for score component } i \text{ to increase} \\ 1 \text{ otherwise} \end{cases} \tag{5}$$

This approach is distinct from the one typically used for *Cost-Benefit Analysis* (CBA) tasks in urban planning practice, in that it seeks to optimize an aggregate function of the relative improvements in each KPI rather than optimizing the net improvement from all KPIs. While CBA often draws skepticism due to the discretion inherent in the process of converting all KPIs into a common unit such as time or money so that the net value of costs and benefits can be computed, the approach taken in the BISTRO scoring function does not require any such assumptions to be made. Rather, each score component represents the relative improvement over the BAU that is achieved by a simulation run using a particular set of UDIs. Objective function designers may choose to apply additional scaling factors to the score components using the $\vec{\alpha}$ parameter vector.

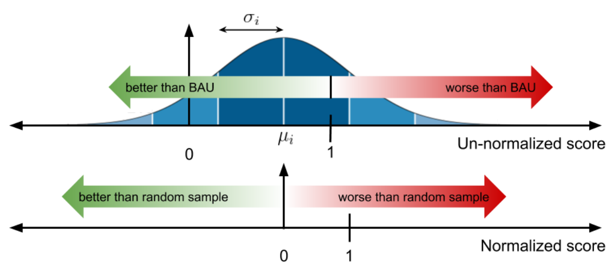

Fig. 3. A visual representation of the normalization procedure for a hypothetical score component, i. The ratio of the submission score and BAU score (depicted in the upper plot) is normalized by taking its z-score (depicted in the lower plot) relative to a random input sample.

3.3 Inputs

Preparation of fixed inputs. For each BISTRO study, a set of fixed inputs must be provided to BEAM. For a given study area, these typically include the road network, the transit schedule, and the demand profile. Depending on the system objectives, additional data may be necessary to fully configure the simulation.

The *road network*, including the physical properties of its links and nodes, may be generated using *Open Street Maps* (OSM) data for the geography of interest. The *transit network* configuration follows the easily accessible *General Transit Feed Specification* (GTFS) format. *On-demand ride services* (*Transportation Network Companies* [TNCs] such as Uber and Lyft), are modeled as a fleet of vehicles driven by agents that are exogenous to the population, or may be driven autonomously. The initial locations of the vehicles may be sampled randomly or from a specified distribution in accordance with appropriate data. The price of on-demand rides is fixed, consisting of a distance-based and a duration-based component. The size of the on-demand ride service fleet is a proportion of the total number of agents in the simulation, as determined by a configuration parameter. Driver repositioning behavior when not currently driving to or serving a passenger can be configured to follow one of several repositioning algorithms defined within BEAM.

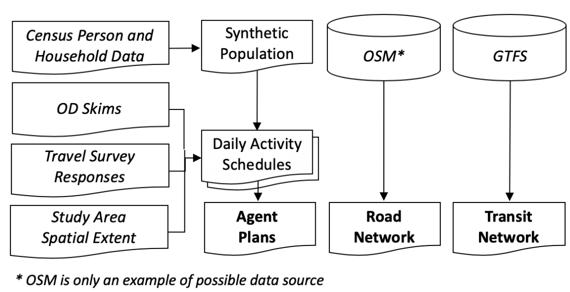

Fig. 4. Generation of fixed inputs. Italicized entities represent the necessary data to generate fixed inputs (in bold).

At the start of the simulation, a *synthetic population* of virtual agents and households is generated such that the sociodemographic attributes of these virtual entities are spatially distributed in accordance with real-world census and/or location-based data from the city of interest. Each agent follows a daily plan consisting of several activities throughout the day. As illustrated in Figure 4, these *daily activity schedules* are generated based on *origin-destination* (OD) skims (matrices that provides the number of trips between zones), travel surveys, and zonal boundary spatial data.

Calibration. Prior to use in optimization runs, the simulation needs to be calibrated to ground truth data. There are multiple possible calibration targets. Calibration is typically targeted at mode split, volumetric traffic counts, and travel distance distributions. Parameters changed during calibration adjust properties that control these target ground truth numbers, *e.g.*, road, transit, or parking capacity scaling factors. Agent behavioral parameters are also sometimes updated as part of this process. The choice of which to use depends on the specific aspects of the city model that are of greatest interest to project planners. We may consider calibration to be an instance of the urban transportation system optimization problem described in Section 2.3, which permits flexible calibration to be achieved significantly faster than manual tuning.

Data resolution and quality can guide calibration constraints, *e.g.*, network data resolution. Due to computational constraints, often a sub-sample of a full population is simulated in an ABMS and then scaled up to the full population for system evaluation, which might also introduce artifacts contributing to the need for calibration.

Configuration of UDIs. BISTRO provides a library of possible inputs for scenario designers to adapt to specific use cases. The selection of UDIs is intended to be compatible with the system objective. UDIs may represent, for example, the investment (*e.g.*, transit fleet mix modification, bus route modifications, parking supply, electric vehicle charge station locations, dynamic redistribution of e-bikes or on-demand vehicles), incentive (*e.g.*, incentives to specific socio-demographic groups for selected transportation modes, road pricing/toll roads, fuel tax), or policy/operational (*e.g.*, transit schedule adjustment, transit fare modification, parking pricing) levers applicable to the study at hand. The project owner may constrain the range of possible values upon which each UDI is valid by setting the corresponding input validation parameters and business rules. The example input file for bus scheduling shown in Table 1 defines alteration of the headway of a particular bus route during a particular service period (defined by its start and end times).

Table 1. Example of bus frequency adjustment input file.

route_id	start_time	end_time	headway_secs	exact_times
1340	21600	79200	900	1
1341	21600	36000	300	1
1341	61200	72000	300	1

3.4 Output Analysis and Visualization

The outputs of calibrated simulations of transportation systems allow stakeholders to better understand the implications of proposals identified using optimization algorithms. For example, visualizations of congested roadways with millions of agents behaving independently can provide a concise method to communicate the effects of infrastructure interventions. To help users with the interpretation of these outputs, the BISTRO platform provides a suite of tools that parse the raw output from BEAM simulation runs and facilitate multi-variate analysis and visualization.

The *parser* processes the raw outputs of a BEAM simulation run into a collection of tables in the structure of a relational database. An easy-to-use Jupyter notebook template is provided that imports six primary table that can be used for complete analysis of the results of any simulation run: `Household`, `Person`, `Activity`, `Trip`, `Leg`, `PathTraversal`

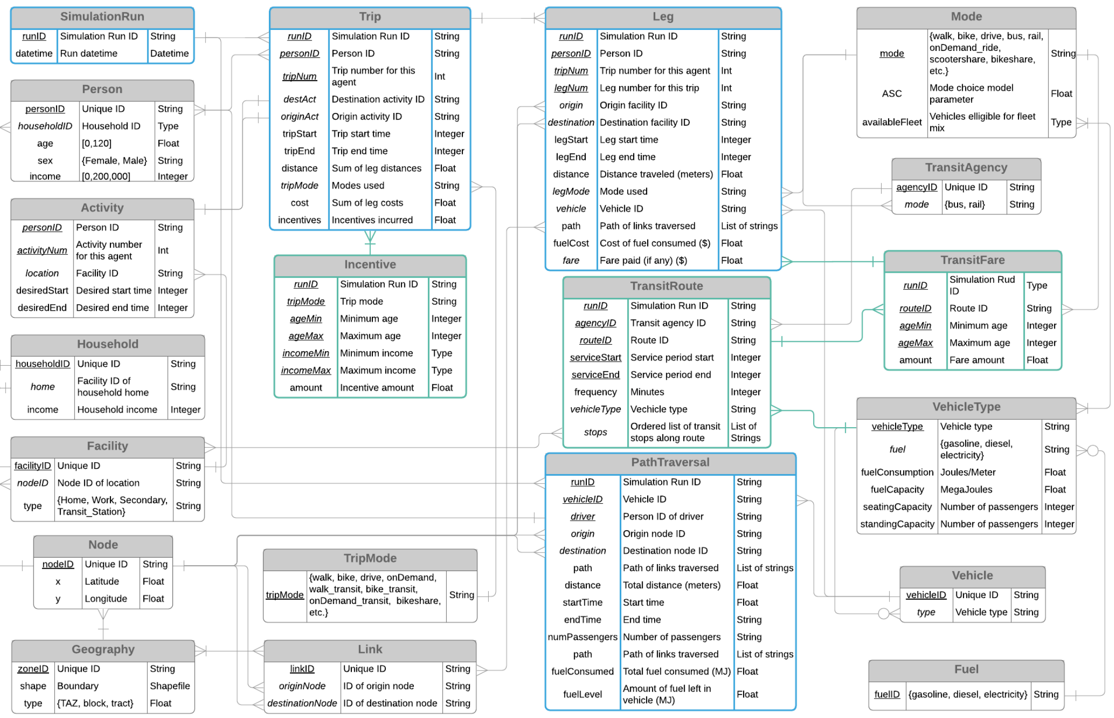

Fig. 5. BISTRO output database entity-relationship diagram. Each entity is identified by a unique primary key (underlined) made up of one or more attributes. Primary keys may include foreign keys (italicized), or attributes that uniquely identify another entity. All entities shown in grey are fixed across simulation runs for a given scenario. Entities highlighted in green are example BISTRO inputs. Entities highlighted in blue are simulation outputs. Thus, the physical entities (`Geography`, `Node`, `Link`, and `Facility`), transportation mode-defining entities (`VehicleType`, `Fuel`, `TransitAgency`, `Mode`, and `TripMode`), and population-defining entities (`Household`, `Person`, and their `Activity`) do not change across simulation runs. Each `SimulationRun` produces unique travel behaviors (`Trip`, `Leg`, and `PathTraversals`). The inputs defined for each `SimulationRun` may influence travel behavior through the pathways highlighted in green.

(see Figure 5). All tables can be linked together by one of two common variables: the `personID` or the `vehicleID`. The `Person` table includes the personal attributes of each agent (*e.g.*, her `householdId`, age, sex and income), identified by her unique `personId`. Similarly, the `Household` table defines the household attributes as well as the home location of the household. The `Activity` table defines the characteristics of each planned activity (*e.g.*, its location, desired start time, duration, desired end time). While the `Person`, `Household`, and `Activity` tables are fixed for a particular scenario, the following three tables (the `Trip`, `Leg`, and `PathTraversal` tables) are generated after each simulation run and thus are identified uniquely by a combination of the simulation run and other key attributes. The `Trip` and `Leg` tables describe agent movements. During the simulation, person agents make one or more `tours` of travel to sequential activities, starting and ending each tour at home. Each `trip` in a tour represents travel from one activity to the next. Trips may consist of one or more `legs` of travel, each using a particular `mode` of transportation. Finally, the `PathTraversal` table describes the vehicle movements and their features (*e.g.*, `driverId`, number of passengers, distance, fuel consumed, etc.).

The *visualization notebook* helps users to understand the concrete impacts of a set of policy inputs on the transportation network. The user can visualize the inputs of the simulation as well as BISTRO KPIs.

3.5 Implementation Details and Performance Characteristics

Both BISTRO and BEAM are primarily implemented in Scala. Input files are read from a single directory and injected into the BEAM initialization routine. The system is containerized using Docker, which helps to facilitate OS-agnostic local and remote execution.

It is anticipated that a run of BEAM on a given set of inputs will be the most compute intensive aspect of a BISTRO optimization routine. While simulation times will vary from scenario to scenario based on model complexity, currently, the primary performance bottleneck in BEAM is routing. The routing engine generates millions of routes (reflecting multimodal options for agents to choose between) for a single simulation run. Some additional overhead considerations such as data availability, level of model resolution required, as well as the impact of augmented BEAM functionality must be balanced in light of available computational resources.

4 INITIAL PILOT STUDY AND LAUNCH

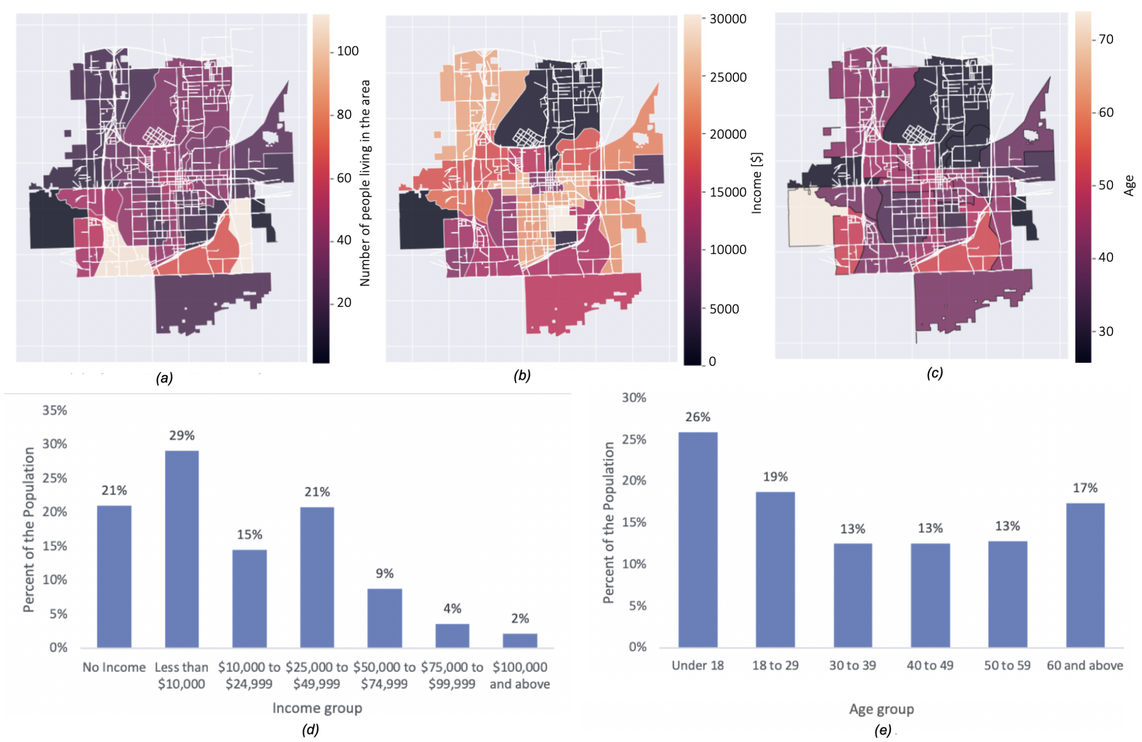

Fig. 6. Demographics of Sioux Faux. (a) Overall distribution of the population per census tract. (b) Distribution of the median population income per census tract. (c) Distribution of the median population age per census tract. (d) Overall population income distribution. (e) Overall population age distribution.

4.1 Sioux Faux

An agent-based model of transport supply and demand inspired by the real city of Sioux Falls, South Dakota[14] was adapted for the purpose of developing and testing example scenarios within BISTRO. To underscore that for these purposes, such scenarios were not developed to be true replicas of the city of Sioux Falls, this benchmark BISTRO scenario is referred to as *Sioux Faux*. The scenario configuration, input specification, and scoring function were designed to investigate the trade-offs between transit service provision, operational costs, mobility, and sustainability.

4.1.1 Scenario Configuration.

Population and plan synthesis. The synthetic population of Sioux Faux was generated using publicly-available survey data for the city of Sioux Falls, South Dakota and the Doppelganger library,[15] a state-of-the-art population synthesis framework developed in Python. Specific inputs to Doppelganger used to generate the Sioux Faux population included household and individual *Public Use Microdata Sample* (PUMS) data for South Dakota from the 2012-2016 (5-year) *American Community Survey* (ACS), which is conducted annually by the US Census.[16] The *Public Use Microdata Area* (PUMA) for Sioux Falls constrains the state-wide survey data to our general area of interest.

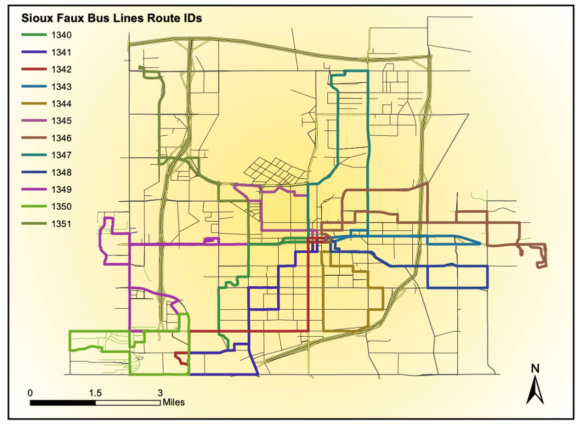

Fig. 7. Sioux Faux bus and road networks.

An existing set of agent plans for Sioux Falls developed for MATSim simulations was used as the basis for the plans of our expanded Sioux Faux population.[17] After initial pilot testing to determine trade-offs between population size, behavioral realism, and computational complexity, we took a 15% sub-sample of the full synthetic population (approximately 15,000 agents). We used a spatially-constrained sampling mechanism in order to allocate plans to agents in accordance with household locations and census tract household and individual attribute distributions. The subsampling mechanism also enforces logical assumptions such as "agents under the age of 18 should not have a work activity" and "agents under the age of 16 should not be allowed to drive".

Transportation Network. The Sioux Faux transportation network includes a road network accessible to walking agents, personal vehicles, on-demand ride services (TNCs such as Uber and Lyft), and public buses providing fixed-route service.[18] The on-demand ride services implemented in this scenario include only single-passenger rides (*e.g.*, UberX, Lyft Classic) from a fleet of on-demand ride vehicles that was distributed randomly across the road network at the start

[14]The "Sioux Falls" scenario is a commonly used benchmark in ABMS research, see https://github.com/bstabler/TransportationNetworks/tree/master/SiouxFalls

[15]For more information about the Doppelganger library, see https://github.com/sidewalklabs/doppelganger

[16]The 5-year PUMS comprises a 5% sample of the US population. It is computed as an aggregate of 1-year samples, which themselves aim to survey 1% of the US population

[17]For more information on the Sioux Falls scenario, see https://www.ethz.ch/content/dam/ethz/special-interest/baug/ivt/ivt-dam/vpl/reports/901-1000/ab978.pdf

[18]The initial bus route scheduling is directly generated from the publicly available GTFS for Sioux Falls, which includes erratic headways across routes.

of each simulation run. The modal split in the BAU configuration of Sioux Faux is heavily dominated by personal car usage with approximately 75% of the miles traveled.

4.1.2 User-Defined Input Specification. For this initial pilot study, a set of four UDIs were investigated: bus fleet vehicle composition, bus service frequency, bus fare, and a multimodal incentive program for on-demand rides and public bus trips. In the BAU bus fleet, all vehicles were set to a default bus type. Optimization of the bus fleet vehicle composition and service frequency offers the opportunity to improve the level of bus service by better matching the bus type with specific demand characteristics of each route. Four types of buses (including the default) were considered (see Table 2), each with different technical properties (seating and standing capacity) and cost characteristics (cost per hour, cost per mile, fuel type and fuel consumption rate).

Table 2. Transit vehicle types available for Sioux Faux bus fleet. (a) Fuel type, (b) Fuel consumption rate (J/m), (c) Operational cost (USD/hr), (d) Seating capacity, (e) Standing capacity.

Vehicle type, $c \in C$	a	b	c	d	e
BUS-DEFAULT	diesel	20048	89.88	37	20
BUS-SMALL-HD	diesel	18043.2	90.18	27	10
BUS-STD-HD	diesel	20048	90.18	35	20
BUS-STD-ART	diesel	26663.84	97.26	54	25

A UDI was implemented to vary the bus schedule on each route, including the hours of service and the headway, or service frequency as shown in Table 1. Multiple service periods with varying headways on the same route were thus possible. The bus fare UDI allowed for the optimization of the fare on each route, segmented by passenger age groups. Finally, a multimodal incentive UDI was implemented to enable reimbursement for on-demand rides, walk to/from transit, or drive to/from transit trips to qualifying individuals based on age, income, or both.

Table 3. Values used for α_i in each of the subsequent results sections. For score components that are positively related to desirable outcomes, negative α_i is provided to transform it consistent with a minimization problem.

KPI	KPI type	Contest	Post-Contest	New KPIs
accessible work locations	Accessibility	-1	-1	–
accessible secondary locations	Accessibility	-1	-1	–
accessible work locations by car	Accessibility	–	–	-1
accessible secondary locations by car	Accessibility	–	–	-1
accessible work locations by transit	Accessibility	–	–	-1
accessible secondary locations by transit	Accessibility	–	–	-1
average trip expenditure-work	LoS	1	1	–
average trip expenditure-secondary	LoS	1	1	–
average travel cost burden-work	Equity	–	–	1
average travel cost burden-secondary	Equity	–	–	1
average bus crowding experienced	LoS	1	1	1
total vehicle miles traveled	Congestion	1	1	1
average vehicle delay per passenger trip	Congestion	1	1	1
costs and benefits	Financial Sustainability	-1	-1*	-1*
total grams $PM_{2.5}$ emitted	Environmental Sustainability	1	1	1
total grams GHG_e emitted	Environmental Sustainability	–	–	1

*fixed KPI post-contest

4.1.3 Business Rules. In order to ensure that optimal solutions would be compliant with common policy and planning practices, four business rules were implemented: 1) there may be no more than five distinct bus service periods (this mimics the typical delineation: am peak, midday, pm peak, evening, late night/early morning), 2) bus route headways may be no more than 120 minutes and no fewer than 3 minutes, 3) bus fares and mode incentives may not isolate a single age, and 4) ages for both fares and incentives may be specified in segments no smaller than five years in range and income for incentives may be assigned in segments no smaller than $5,000 in range.

4.1.4 Scoring Function Design. The set of Sioux Faux UDIs have varying interconnected impacts on the operation of and access to public transit and on-demand ride service by agents. Thus, the scoring function upon which the inputs were optimized was designed to include a variety of metrics that relay feedback on the user experience and operational efficiency of the transportation system as a whole.

Five user experience KPIs were developed to represent three main aspects of mobility: accessibility, travel expenditure, and transit passenger comfort. The accessibility and travel expenditure were both disaggregated by trip purpose such that score components for accessibility and travel expenditure to work and secondary activities were each included separately in the scoring function. Transit passenger comfort was measured as the average bus crowding experienced by bus passengers.[19]

Four KPIs of operational efficiency were included to account for the congestion, environmental sustainability, and financial sustainability resulting from optimized inputs. Total VMT was included as a KPI for overall congestion while average vehicle delay per passenger trip served as a KPI of the average impact of congestion. The total amount of $PM_{2.5}$ emitted served as a KPI of the environmental impact resulting from each simulation run. Finally, the financial sustainability KPI was included to incentivize outcomes with minimal impact to the bottom line of the transit agency by taking into account the operational costs, incentives distributed, and revenues collected from any combination of transit fleet mix, scheduling, fare structure and incentive program.

A random input sample of 800 runs produced the normalizing parameters for each metric. All metrics are aggregated according to the following function:

$$F\left(\vec{C}_a, \vec{F}, \vec{\sigma}, \vec{\mu}, \vec{\alpha}\right) = \sum_{i \in \mathbb{K}} \frac{\left(\frac{K_i(C_s)}{K_i(C_{BAU})}\right)^{\alpha_i} - \mu_i}{\sigma_i} \tag{6}$$

where all variables are defined as described in Section 3.2.2, with the set of KPIs and corresponding α_i values as specified in Table 3.

4.2 Results of internal testing for over 400 users

Contest participation. Over the course of 17 days, 487 people in teams of one to four (mostly consisting of engineers and data scientists with little to no domain expertise in transportation planning) effectively created nearly 1,000 different "city transportation plans."[20] While these plans lacked the necessary public policy context to be considered as implementable solutions, this case study represents the first time that an exercise in transportation system optimization was conducted at the scale of hundreds of researchers and demonstrated that BISTRO can enable transportation policy scenario planning at scale.

To be able to compare their results and scores with other participants, each team could submit up to five solutions per day and thus be ranked in a web-accessible leaderboard. While contestants trained algorithms online, final evaluation, leaderboard, and discussion boards were hosted by AICrowd.com.[21] Inputs from top teams were evaluated 5 times for 100 iterations each in order to achieve a consistent final score. Figure 8 illustrates the evolution of submissions over time during the competition. Participation developed in two phases. During the first week, contestants became familiar with the BISTRO framework and the Sioux Faux transportation optimization problem. During the second phase, contestants continued to optimize their solutions.

[19] Average bus crowding in the Sioux Faux scenario was calculated as the average number of agent hours spent per transit trip in buses occupied above their seating capacity. This KPI has since been updated, see Section 3.2.1

[20] Uber does not endorse any of the solutions presented.

[21] http://www.aicrowd.com

According to code submissions and a post-contest survey, the teams with the best solutions followed similar strategies. Typically, they used domain-specific analysis to prune the large input space. For instance, some teams decided to apply the same fare structure on all bus lines with a reduced fare for people under 25 years of age and a full fare for all older riders. Thus, they effectively reduced the number of settings for the bus fare input from $2^{12} = 4096$ to $2^{1} = 2$ combinations. Contestants then used a variety of algorithmic approaches to automatically search the reduced design space. As shown in Table 4, the black-box global optimization techniques used during the pilot study primarily incorporated variants of Bayesian optimization, genetic/evolutionary algorithms, gradient-based techniques, and meta-heuristics methods.

Most teams managed to improve their scores by five standard deviations better than a random search benchmark (*i.e.*, with scores of approximately -5). Due to an insidious modeling deficiency, the financial sustainability score component could be optimized towards negative infinity. As such, any contributions from other score components would be relatively inconsequential. Two teams discovered input settings that took advantage of the lack of a lower bound on the financial sustainability score component and were thus able to reach extremely low scores of -30 or -40. This experience underlines the importance of developing careful theory supplemented by judicious testing when designing objective functions.

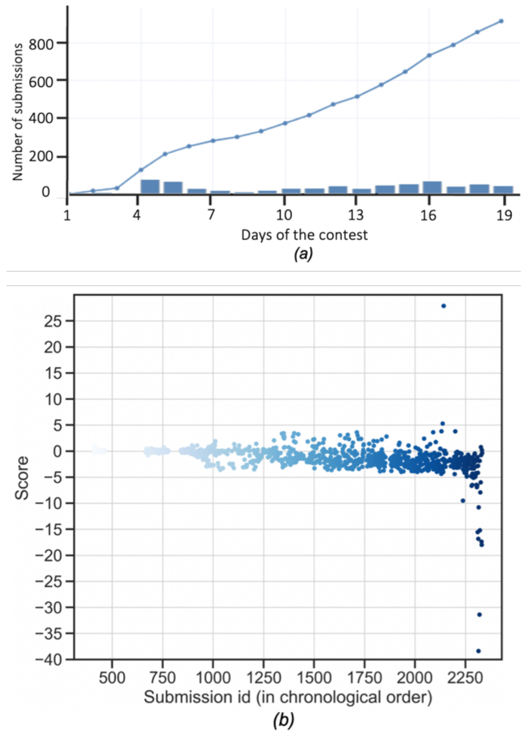

Fig. 8. Participation history. (a) Number of solutions submitted over time (per day and cumulative). (b) Evolution of scores over time.

Post-contest studies results. After the contest, two follow-on studies were conducted to interpret the solutions from the top algorithms in the context of improvements to the objective function. An initial set of improvements, hereby referred to in this text as the "Post-Contest" objective function study, simply addressed the unbounded financial sustainability score component as well as other minor problems discovered during the competition. The "New KPIs" objective refers to an expanded set of KPIs, summarized in Table 3. Two of the best-performing algorithms from the competition—namely, Bayesian Optimization using *tree-based Parzen estimators* (TPE) [42] and *Genetic Algorithms* (GA) [54]—were adapted and re-implemented to run BISTRO on the 15k Sioux Faux scenario with the updated objective functions. The GA assessment on the "New KPIs" objective utilized five parallel evolutions, each drawing a random sample

Table 4. Proportion of algorithmic approaches used, according to a survey conducted with contestant teams.

Approach	Proportion
Bayesian optimization	34%
Evolutionary algorithms	28%
Gradient based	14%
Meta-heuristics	7%
Plackett-Burman design	3%
Hill climbing	3%
Other	10%

of seeds from a larger gene pool. As a baseline benchmark, *random search* (RS) was performed for 800 trials. Both TPE and GA were run over similar design spaces for 1,400 trials on the “Post-Contest” and “New KPIs” studies.[22]

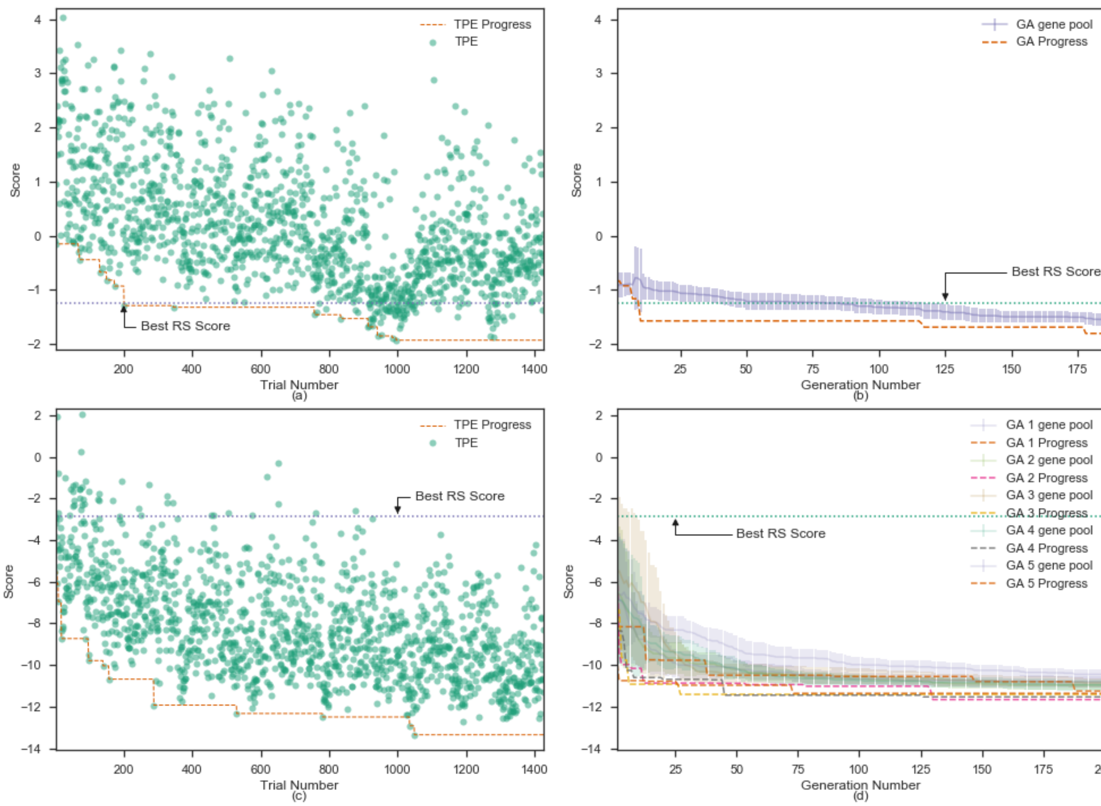

Fig. 9. Optimization of the Sioux Faux 15k scenario with TPE (left) and GA (right) using “Post-Contest” (top) and “New KPIs” (bottom) objective function settings. The dashed line(s) across the bottom of each denotes the best (lowest) score achieved by an algorithm within the first N trials. Individual trial scores (at 40 iterations) are shown for TPE plots, whereas one standard deviation ranges of current gene pools are displayed in the GA plots. For the “Post-Contest” objective, the TPE and GA algorithms surpass the best score from 800 RS trials of 40 iterations (-1.24) within 200 and 10 trials, respectively. For the “New KPIs” objective, both algorithms significantly outperform the best result (-2.84) of an 800 trial, 40 iteration RS almost immediately.

Inputs corresponding to the trial yielding the top score for each algorithm were then simulated for 100 iterations with five replicates. Figure 9 demonstrates that both GA and TPE produce input configurations that are superior to RS. Notably, both GA and TPE achieve solutions that reflect sensible yet distinct transportation system management strategies. Both solutions trade off financial sustainability for other KPIs. However, the two solutions achieve their results via distinct means. The TPE solution runs its buses with medium-to-high fares and medium-to-high incentives across all lines with diverse frequency change choices throughout the day. Conversely, the GA solution offers free

[22]Partial convergence criteria of 40 iterations were used during initial search, as this was determined to be sufficient for establishing a trajectory consistent with a fully relaxed state.

transit (for “Post-Contest”) and mostly free transit (for “New KPIs”, only charging the 16-65 age group $2-$3 on a few bus lines) and does not specify any incentive strategies. Figure 10 suggests that differences in TPE and GA solutions for the “Post-Contest” objective arise from distinct transit usage patterns. The TPE solution tends to optimize the bus LoS around work and secondary activity start times, whereas the GA solution ensures that buses are available during the evening peak commute time.

Figure 11 presents visualizations of input distributions for the top fifth percentile of TPE trials. It is evident from this figure that the UDI values for the best performing (lowest scoring) solutions occupy a narrow band in the design space. Arriving at tight distributions of inputs is indicative of algorithm convergence as well as objective function sensitivity to input settings. For example, the best performing TPE input sets evaluated in BISTRO under the “Post-Contest” objective suggest charging higher bus fares for adult citizens (16-60) than for youth (1-15) and elderly (60-120). The corresponding bus types on a given route suggested by these solutions are well-resolved and tend towards smaller models. In contrast, for near-optimal inputs evaluated using the “New KPIs” objective function, fares assigned to youth are, on average, higher than those assigned to adults and seniors. The corresponding bus types by route are also more diverse among optimal solutions, indicating that the objective is less sensitive to the `VehicleFleetMix` input when evaluated using the “New KPI” objective function. Using the “New KPIs” objective, GA (not shown) also finds a tight distribution of fares for top-performing solutions but contrarily finds diversity in its `VehicleFleetMix` solutions.

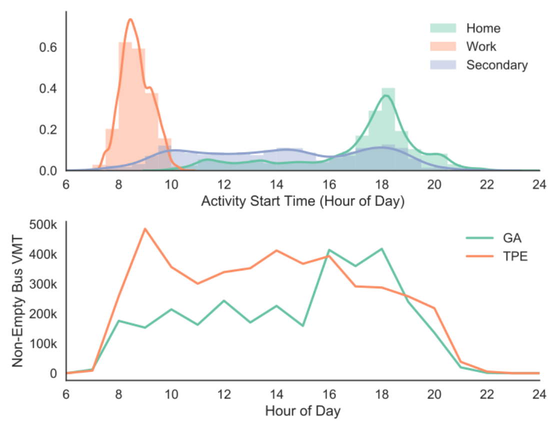

Fig. 10. Example of output analysis for the “Post-Contest” case study. The upper plot shows the various activity start times of agents by activity type. The lower plot shows bus crowding (*i.e.*, ridership exceeding half of seating capacity) for two competing algorithms.

5 CONCLUSION

BISTRO represents the first general-purpose transportation policy decision support tool and scenario-based optimization framework supported by empirically-driven agent-based models. When combined with sensible guidance from experienced planning professionals, BISTRO can be used to identify more holistic, empirically-driven approaches to urban transportation planning and management. BISTRO does this by encouraging an online, collaborative process through which all stakeholders can develop system objectives, identify potential policy and/or investment opportunities, and visualize trade-offs identified during search for optimal combinations of proposed interventions.

In addition to overall system purpose, design, and software architecture, this work provides a concrete example of the process that BISTRO supports as implemented in the context of a scenario-based policy optimization “contest.” Participants were tasked with identifying operational investments, fare pricing strategies, route scheduling modifications, and travel mode-specific incentives that, when applied to a toy replica of daily travel in Sioux Falls, SD, yielded the best value of an objective function aggregating score components that measured improvements in various KPIs of transportation system health over the *status quo*. While many participants had little or no prior expertise in the transportation science and policy analysis methods typically used in urban planning practice, over a dozen teams

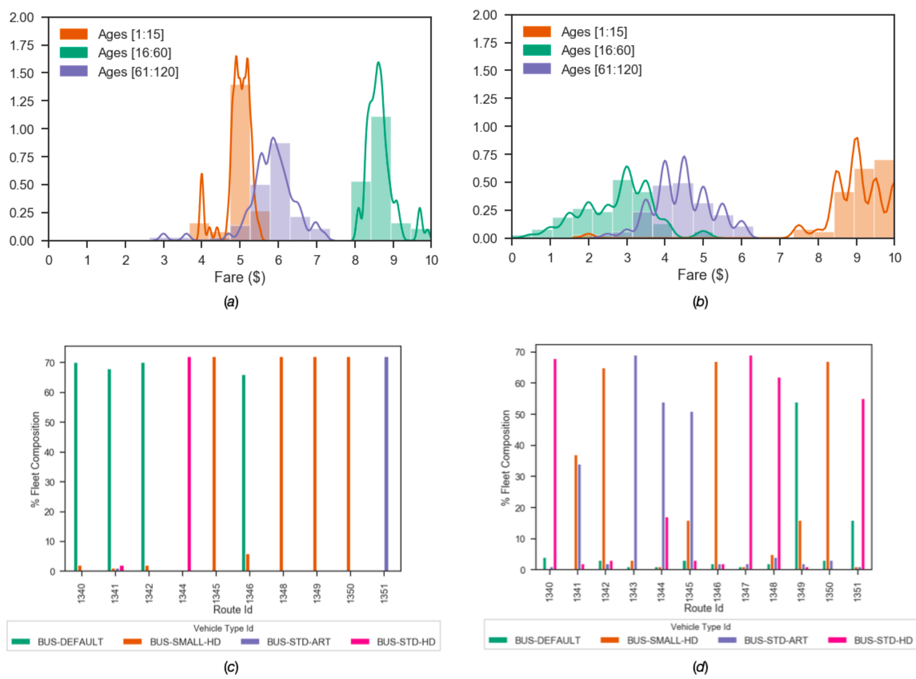

Fig. 11. Distributions of bus fare by age (top) and vehicle fleet mix by route (bottom) for inputs representing the best fifth percentile scores among trials run for Sioux Faux 15k scenario using the TPE algorithm, shown for “Post-Contest” (left) and “New KPIs” (right) objective functions.

developed algorithms that found inputs, which, when evaluated in the simulator, achieved scores that surpassed both random search as well as human judgement. A visualization toolkit was developed in order to provide stakeholders with a better understanding of how surprising changes in aggregate transportation system statistics emerge due to the complex effect of selected intervention strategies on the microscopic decision-making behavior of thousands of agents.

Despite rigorous analysis and testing prior to the release of the pilot study, the two top scoring teams discovered that one of the KPIs had been inadvertently defined over an unbounded region of the search space, resulting in winning entries that provided little, if any, value from a decision support perspective. The mixed results of the competition led us to conclude that the optimization-based search techniques enabled by BISTRO should support an iterative approach that involves applying optimization algorithms to refinements of KPI specifications in order to better align objective functions with system goals. Towards this end, we adapted algorithms from the two winning teams to optimize two revisions of the KPIs comprising the Sioux Faux objective function.

Research conducted using BISTRO strives to meet the highest standard of reproducibility in computational experiments [55–57] as well as fact-based policymaking [58] by making all data, models, and algorithms freely available and open source[23]. Consequently, in contrast to commercial (as well as many publicly available) urban planning DSS tools, BISTRO increases transparency in public decision-making while improving the robustness of experimental findings. One finding of post-contest reproducibility efforts was that different classes of algorithms appeared to converge to

[23] All code and data used and referred in the article is available at [URL-TBA]

solutions that emphasized distinct policy strategies, implying that BISTRO is amenable to multi-objective problem formulations and algorithmic approaches.

This conclusion suggests that, in addition to its utility as a decision support system, BISTRO could serve as an exemplary testbed for multiple emerging streams of research (*e.g.*, freeze-thaw, multi-objective, multi-task, and multi-fidelity optimization) in SMBO and associated meta-model-based optimization methods. Should BISTRO be widely adopted as part of the urban planning toolkit, innovative algorithms and new theory developed as part of inquiry in these sub-domains will have the added benefit of directly serving a humanitarian purpose. In effect, BISTRO leverages the distinct backgrounds of planners and computer scientists to facilitate a process to draw upon the strengths of two complementary areas of expertise to *inform* rather than *direct* public conversations about proposed policy, investments, and regulations.

ACKNOWLEDGMENTS

The authors wish to thank the Uber employees who participated in the contest and provided valuable feedback.

REFERENCES

[1] Mogeng Yin, Ziheng Lin, Sid Feygin, Madeline Sheehan, and Jean-Francois Paiment. AM/PM: Travel Demand Nowcasting. 2018.

[2] Hong Zheng. A Primer for Agent-Based Simulation and Modeling in Transportation Applications. *Development*, page 75, 2013.

[3] USDOT. Trends in Statewide Long-Range Transportation Plans: Core and Emerging Topics, 2012.

[4] Federal Highway Administration/Federal Transit Administration. Public Engagement: Case Studies and Notable Practices, 2019.

[5] Gian-Claudia Sciara. Metropolitan transportation planning: Lessons from the past, institutions for the future. *Journal of the American Planning Association*, 83(3):262–276, 2017.

[6] Metropolitan Transportation Commission. Legal Settlements, 2019.

[7] Brett Simpson. CP&DR News Briefs March 5, 2019: SANDAG Transportation Plan; SoMa Lawsuit; Housing Shortfalls Statewide; and More, 2019.

[8] Joe Linton. 'Fix the City' Lawsuit Challenges L.A.'s Expo Line Development Plan, 2018.

[9] Joshua Auld, Michael Hope, Hubert Ley, Vadim Sokolov, Bo Xu, and Kuilin Zhang. Polaris: Agent-based modeling framework development and implementation for integrated travel demand and network and operations simulations. *Transportation Research Part C: Emerging Technologies*, 64:101–116, 2016.

[10] Carlos Lima Azevedo Moshe Ben-Akiva Jessika Transik. TRIPOD: Sustainable Travel Incentives with Prediction, Optimization and Personalization, 2019.

[11] Ebru Vesile Ocalir-Akunal. Decision Support Systems in Transport Planning. *Procedia Engineering*, 161:1119–1126, 2016.

[12] The transportation planning process key issues: A briefing book for transportation decisionmakers, officials, and staff. Technical report, Washington, DC, USA, 2016.

[13] Report to the ranking member, committee on environment and public works, u.s. senate– metropolitan planning organizations: Options exist to enhance transportation planning capacity and federal oversight. Technical report, Washington, DC, USA, 2009.

[14] Joe Castiglione, Mark Bradley, and John Gliebe. *Activity-Based Travel Demand Models: A Primer.* 2016.

[15] D Helbing and S Balietti. How to do agent-based simulations in the future: From modeling social mechanisms to emergent phenomena and interactive systems design (ssrn scholarly paper no. id 2339770). *Rochester, NY: Social Science Research Network. Retrieved from the Santa Fe Institute Archive: http://www. santafe. edu/research/workingpapers/abstract/51b331dfecab44d50dc35fed2c6bbd7b*, 2013.

[16] Alison Heppenstall, Nick Malleson, and Andrew Crooks. "Space, the Final Frontier": How Good are Agent-Based Models at Simulating Individuals and Space in Cities? *Systems*, 4(1):9, 2016.

[17] L. Smith, R. Beckman, and K. Baggerly. TRANSIMS: Transportation analysis and simulation system. In *Fifth National Conference on Transportation Planning Methods Applications-Volume II: A Compendium of Papers Based on a Conference Held in Seattle, Washington in April 1995Transportation Research Boad and Washington State Department of Transportation*, 1995.

[18] Kay W. and Axhausen. The Multi-Agent Transport Simulation MATSim. *The Multi-Agent Transport Simulation MATSim*, 2016.

[19] Michael Behrisch, Laura Bieker, Jakob Erdmann, and Daniel Krajzewicz. Sumo–simulation of urban mobility: an overview. In *Proceedings of SIMUL 2011, The Third International Conference on Advances in System Simulation*. ThinkMind, 2011.

[20] Carlo Castagnari, Flavio Corradini, Francesco De Angelis, Jacopo de Berardinis, Giorgio Forcina, and Andrea Polini. Tangramob: an agent-based simulation framework for validating urban smart mobility solutions. *arXiv preprint arXiv:1805.10906*, 2018.

[21] Joseph Kamel, Reza Vosooghi, Jakob Puchinger, Feirouz Ksontini, and Göknur Sirin. Exploring the Impact of User Preferences on Shared Autonomous Vehicle Modal Split: A Multi-Agent Simulation Approach. *Transportation Research Procedia*, 37(September 2018):115–122, 2019.

[22] Paul Salvini and Eric J Miller. Ilute: An operational prototype of a comprehensive microsimulation model of urban systems. *Networks and spatial economics*, 5(2):217–234, 2005.

[23] Paul Waddell. Urbansim: Modeling urban development for land use, transportation, and environmental planning. *Journal of the American planning association*, 68(3):297–314, 2002.

[24] T Nicolai and K Nagel. Coupling transport and land use: Investigating accessibility indicators for feedback from a travel to a land-use model. In *Latsis Symposium*, pages 12–16, 2012.

[25] Rolf Moeckel, Carlos Llorca Garcia, Ana Tsui Moreno Chou, and Matthew Bediako Okrah. Trends in integrated land use/transport modeling: An evaluation of the state of the art. *Journal of Transport and Land Use*, 11(1):463–476, 2018.

[26] Thomas W Nicolai and Kai Nagel. Coupling matsim and urbansim: Software design issues. Technical report, SustainCity Working Paper, 2010.

[27] Dominik Ziemke, Kai Nagel, and Rolf Moeckel. Towards an agent-based, integrated land-use transport modeling system. *Procedia computer science*, 83:958–963, 2016.

[28] US Department of Energy. Energy Efficient Mobility Systems: 2018 Annual Progress Report. Technical report, US Department of Energy, Office of Energy Efficiency and Renewable Energy, Vehicle Technologies Office, dec 2018.

[29] Daniel Delling, Thomas Pajor, and Renato F Werneck. Round-based public transit routing. *Transportation Science*, 49(3):591–604, 2014.

[30] Matthew Wigginton Conway, Andrew Byrd, and Marco van der Linden. Evidence-Based Transit and Land Use Sketch Planning Using Interactive Accessibility Methods on Combined Schedule and Headway-Based Networks. *Transportation Research Record: Journal of the Transportation Research Board*, 2653(1):45–53, 2017.

[31] Kenneth E Train. *Discrete choice methods with simulation*. Cambridge university press, 2009.

[32] Moshe E Ben-Akiva, Steven R Lerman, and Steven R Lerman. *Discrete choice analysis: theory and application to travel demand*, volume 9. MIT press, 1985.

[33] GA Agha, Prasannaa Thati, and Reza Ziaei. Actors: a model for reasoning about open distributed systems. *Formal methods for distributed processing: a …*, pages 155 – 176, 2001.

[34] Carolina Osorio and Michel Bierlaire. A simulation-based optimization framework for urban transportation problems. *Operations Research*, 61(6):1333–1345, 2013.

[35] Andrew R Conn, Katya Scheinberg, and Luis N Vicente. *Introduction to derivative-free optimization*, volume 8. Siam, 2009.

[36] Russell R. Barton and Martin Meckesheimer. Chapter 18 Metamodel-Based Simulation Optimization. *Handbooks in Operations Research and Management Science*, 13(C):535–574, 2006.

[37] Carl Edward Rasmussen. Gaussian processes in machine learning. In *Summer School on Machine Learning*, pages 63–71. Springer, 2003.

[38] Bobak Shahriari, Kevin Swersky, Ziyu Wang, Ryan P Adams, and Nando De Freitas. Taking the human out of the loop: A review of bayesian optimization. *Proceedings of the IEEE*, 104(1):148–175, 2015.

[39] Ian Dewancker, Michael McCourt, and Scott Clark. Bayesian Optimization Primer. 2015.

[40] Peter I Frazier. A tutorial on bayesian optimization. *arXiv preprint arXiv:1807.02811*, 2018.

[41] Frank Hutter, Holger H. Hoos, and Kevin Leyton-Brown. Sequential model-based optimization for general algorithm configuration. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 6683 LNCS:507–523, 2011.

[42] JS Bergstra and R Bardenet. Algorithms for hyper-parameter optimization. *NIPS*, 2011.

[43] James Bergstra, Daniel L K Yamins, and D Cox. Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. In *Proceedings of the 30th International Conference on Machine Learning*, pages 115–123, 2013.

[44] Donald R Jones, Matthias Schonlau, and William J Welch. Efficient global optimization of expensive black-box functions. *Journal of Global optimization*, 13(4):455–492, 1998.

[45] Peter Frazier, Warren Powell, and Savas Dayanik. The knowledge-gradient policy for correlated normal beliefs. *INFORMS journal on Computing*, 21(4):599–613, 2009.

[46] Jian Wu, Matthias Poloczek, Andrew G Wilson, and Peter Frazier. Bayesian optimization with gradients. In *Advances in Neural Information Processing Systems*, pages 5267–5278, 2017.

[47] Jasper Snoek, Hugo Larochelle, and Ryan Adams. Practical Bayesian Optimization of Machine Learning Algorithms. In *Advances in neural information processing systems*, pages 2951—-2959, 2012.

[48] Kevin Swersky, Jasper Snoek, and Ryan Prescott Adams. Freeze-thaw Bayesian optimization. *arXiv preprint arXiv:1406.3896*, 2014.

[49] Carolina Osorio and Linsen Chong. A Computationally Efficient Simulation-Based Optimization Algorithm for Large-Scale Urban Transportation Problems. *Transportation Science*, 49(3):623–636, 2015.

[50] Carolina Osorio and Kanchana Nanduri. Urban transportation emissions mitigation: Coupling high-resolution vehicular emissions and traffic models for traffic signal optimization. *Transportation Research Part B: Methodological*, 81:520–538, 2015.

[51] Linsen Chong and Carolina Osorio. A simulation-based optimization algorithm for dynamic large-scale urban transportation problems. *Transportation Science*, 52(3):637–656, 2017.

[52] Ennio Cascetta. A stochastic process approach to the analysis of temporal dynamics in transportation networks. *Transportation Research Part B*, 23(1):1–17, 1989.

[53] Gunnar Flötteröd. A search acceleration method for optimization problems with transport simulation constraints. *Transportation Research Part B: Methodological*, 98:1339–1351, 2017.

[54] David E Goldberg and John H Holland. Genetic algorithms and machine learning. *Machine learning*, 3(2):95–99, 1988.
[55] Roger D Peng. Reproducible research in computational science. *Science*, 334(6060):1226–1227, 2011.
[56] Andrew Morin, Jennifer Urban, Paul D Adams, Ian Foster, Andrej Sali, David Baker, and Piotr Sliz. Shining light into black boxes. *Science*, 336(6078):159–160, 2012.
[57] Geir Kjetil Sandve, Anton Nekrutenko, James Taylor, and Eivind Hovig. Ten simple rules for reproducible computational research, 2013.
[58] Justine S. Hastings, Mark Howison, Ted Lawless, John Ucles, and Preston White. Unlocking data to improve public policy. 03 2019.