Zen: LSTM-based generation of individual spatiotemporal cellular traffic with interactions

Anne Josiane Kouam, Aline Carneiro Viana, Alain Tchana

▶ **To cite this version:**

Anne Josiane Kouam, Aline Carneiro Viana, Alain Tchana. Zen: LSTM-based generation of individual spatiotemporal cellular traffic with interactions. 2023. hal-03910141

HAL Id: hal-03910141
https://inria.hal.science/hal-03910141v1

Preprint submitted on 2 Jan 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Zen: LSTM-based generation of individual spatiotemporal cellular traffic with interactions

Anne Josiane Kouam
Inria
France

Aline Carneiro Viana
Inria
France

Alain Tchana
Grenoble INP
France

ABSTRACT

Domain-wide recognized by their high value in human presence and activity studies, cellular network datasets (i.e., Charging Data Records, named CdRs), however, present accessibility, usability, and privacy issues, restricting their exploitation and research reproducibility. This paper tackles such challenges by modeling Cdrs that fulfill real-world data attributes. Our designed framework, named *Zen* follows a four-fold methodology related to (i) the LTSM-based modeling of users' traffic behavior, (ii) the realistic and flexible emulation of spatiotemporal mobility behavior, (iii) the structure of lifelike cellular network infrastructure and social interactions, and (iv) the combination of the three previous modules into realistic Cdrs traces with an individual basis, realistically. Results show that *Zen*'s first and third models accurately capture individual and global distributions of a fully anonymized real-world Cdrs dataset, while the second model is consistent with the literature's revealed features in human mobility. Finally, we validate *Zen* Cdrs ability of reproducing daily cellular behaviors of the urban population and its usefulness in practical networking applications such as dynamic population tracing, Radio Access Network's power savings, and anomaly detection as compared to real-world CdRs.

CCS CONCEPTS

• **Data and Communication Traffic** → **Charging Data Records**; • **Cellular Traffic** → *Mobility and Network events*; • **Modeling** → **LSTM**.

KEYWORDS

Human mobility modeling, Data and Communication traffic modeling, Recurrent Neural Networks.

1 INTRODUCTION

Charging Data Records are acknowledged as a common tool for studying human mobility, infrastructure usage, and traffic behavior [18]. We name such datasets as CdRs to distinguish them from the standard Call Detail Records (CDRs), describing only call and SMS cellular communication information. CdRs describe time-stamped and geo-referenced event types (i.e., data, calls, SMS) generated by each mobile device interacting with operator networks (cf. Table 1). They comprise city-, region-, or country-wide areas and usually cover long periods (months or years); no other technology currently provides an equivalent per-device precise scope. As a result, CdRs are exploited in different research domains and industries, such as sociology [39], epidemiology [7], transportation [38], and networking [35]. For a quantitative appreciation of such CdRs' worth recognition, Fig. 1 identifies 14 different research domains leveraging CdRs among 100 most relevant works (sorted by Google Scholar) selected from 1022 last 5-year publications. This clearly shows a great diversity of domains in this sample only (~ 10%).

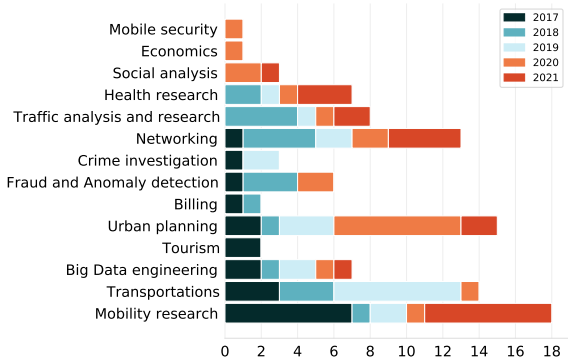


Figure 1: Distribution by domain of the last 5-year most relevant publications using CdRs.

Yet, the exploitation of real-world CdRs for research faces many limitations. First, *accessibility*: CdRs datasets are not publicly available, imposing strict mobile operators' agreements. Second, *usability*: CdRs are usually available in an aggregated form (i.e., grouped mobility flows and coarse spatiotemporal information), limiting related analyses' preciseness. Third, *privacy*: even anonymized, non-aggregated CdRs describe sensitive information of users' habits, which hardens their shareability [28]. This paper *addresses such limitations by enabling the autonomous generation of realistic and privacy-compliant CdRs by scientific community, thus providing new avenues for research advances.*

Moreover, generated CdRs should conform to essential attributes, namely, *completeness, realisticness, fine-grained description, and privacy.* Unfortunately, those attributes make the generation of realistic CdRs challenging and complex. In particular, achieving completeness requires (i) either real-world complete CdRs datasets (hard to obtain) describing mobility, traffic, and pairwise users communications or (ii) to cope with the difficulty in modeling the intrinsic correlations between information describing users' behaviors in space, time, and social communication. Achieving *realisticness* implies considering real-world cellular network complexities (architecture and topology) at all levels of the generation process. The *fine-grained description* achievement is impeded by the heterogeneity of users' behaviors, especially in cellular traffic. Finally, generated traces should be *privacy-compliant* to avoid backtracking real users' identities, most often done through their mobility information.

To the best of our knowledge, *this is the first work in literature producing realistic Charging Data Records (CdRs) that fulfill the*

above-mentioned attributes. Our designed framework, named *Zen* employs a four-fold methodology:

(1) Leveraging on a real-world fully anonymized CdRs describing users' traffic behavior (i.e, events information on its type – data, call, and SMS –, duration, pairwise information, etc), *we propose the first literature modeling that captures long-range and inter-CdRs specificity* correlations while addressing the population heterogeneity (§3). Our model captures population diverseness in the reproduction of individual traffic behaviors. We use three separate *Long-Short-Term Memory neural networks* (LSTM) to model event types generation (i.e., *what*), the inter-event duration (i.e., *when*), the social interactions (i.e., *whom*), and leverage statistical analysis to model CdRs metrics such as calls duration (*i.e., how*). Overall, *Zen* traffic modeling presents significant high performance values, providing for 80% of users (i) more than 95% (for event-type) and 75% (for inter-event) of modeling accuracy, and (ii) less than 6.68% (for inter-event) and 12.5% (for social) of Mean Absolute Error's maximum values.

(2) Mobility behaviors of individuals (§4) are emulated according to the infrastructure of a real-world metropolitan city (here, the Helsinki EU city), and resulting trajectories are coupled with the corresponding cell towers distribution of existing operator networks in the same city [33]. Here, we leverage city planning, transportation information [17] as well literature investigations on laws dictating human mobility [1, 11, 30]. Such real-world information and realistic human mobility modeling are then incorporated in the literature *Working Day Mobility* (WDM) model [9] – extensively enhancing it – to emulate urban daily-life mobility behaviors of individuals. Moreover, we rely on the ONE simulator [20] to bring flexibility to our model regarding population size, duration, and covered area.

(3) We then design a separate module (§5) to realistically reproduce on top of generated mobility traces, a cellular network organization with multiple operators and build social ties between users. This enables the first-of-a-kind flexibility to produce CdRs of numerous operators at the same period.

(4) We combine all the previous models to generate complete CdRs describing individual mobility, traffic, pairwise communications following real traffic behavior (§6).

Note that the disjoint behaviors modeling of realistic emulated mobility and of real-world traffic hides real individuals' spatiotemporal daily-life habits in routine and leisure times (e.g., home/work, nightlife, etc.), bringing the privacy-preserving capability to the produced CdRs.

2 ZEN OVERVIEW

In the following, we provide an overview of *Zen* architecture and describe the different real-world datasets we leverage.

2.1 Architecture

According to input parameters, we generate realistic CdRs (cf. Table 1) through four phases, each implemented in a module of the *Zen* architecture (cf. Fig. 2). *Zen* architecture consists of (1) a *traffic module*, (2) a *mobility module*; (3) a *social-ties module*, and (4) a *CdR-combiner*, or merger module.

The *traffic module* (§3) leverages *Long-Short-Term Memory neural networks (LSTM)* jointly with statistical analysis to model users' traffic behavior from real-world CdRs. It provides answers to *what*, *when*, with *whom*, and *how* to generate events. The *mobility module* (§4) (i) emulates users temporal displacements on a real-world geographical map over a selected period, and (ii) associates corresponding users positions with a real-world cellular topology. This dataset feeds the *social-ties module* (§5) that builds the network social structure on top of which users' communication interactions occur by building users' phonebooks, i.e., list of phone numbers a user is likely to contact. Finally, the *CdR-combiner* module (§6) combines the previous modules' outputs to generate realistic CdRs per network operator over a specified duration and particular urban area.

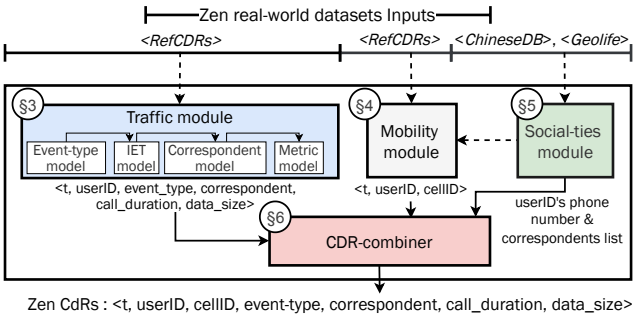


Figure 2: *Zen* architecture.

2.2 Real-world reference datasets

Zen models real-world datasets to produce realistic outputs. In particular, as depicted on top of Fig. 2, *Zen* uses three real-world reference datasets described in what follows.

RefCdRs are used by both the *traffic* and the *social-ties* modules. *RefCdRs* refer to a fully-anonymized CdRs dataset collected by a major mobile network operator. They describe 1-month (*from 2018-06-01 to 2018-06-30*) per-user traffic resulting in about 3 million timestamped events generated by 186,738 distinct phone numbers, where about 17,000 are from the *RefCdRs*' operator. *RefCdRs* are incomplete; they lack mobility features and incoming-SMS traffic type (i.e., only have outgoing SMS). Still, there is no information on the size of sessions in the data traffic type. *RefCdRs* provide each user's operator network code. We leverage this information to identify the list of operators appearing in the datasets.
On the other hand, the *mobility module* leverages the *ChineseDB* [1] and *Geolife* [48] datasets, by extracting statistics describing real-life mobility behavior of users. *ChineseDB* (non-public and fully anonymized mobility CdRs) contains trajectories of 642K users during two weeks. In particular, we did not have access to *ChineseDB* but only to related statistics available in [1]. *Geolife* (public and anonymized GPS dataset) contains trajectories of 182 users during 64 months.

2.3 *Zen* CdRs attributes

We present hereafter the positioning of *Zen* generated CdRs with respect to our goals:

Completeness: Complete CdRs comprise mobility and traffic features and should, thus, include, in addition to user positions (i.e.,

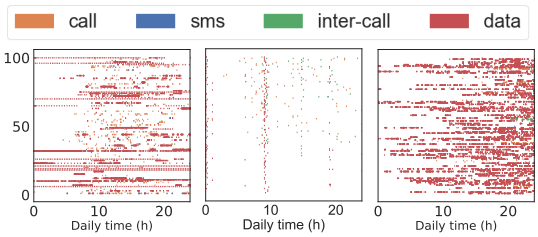


Figure 3: Temporal event sequences of 100 users for: (left) a real-world, (center) a statistically- and (right) the *Zen*-generated CdRs.

Table 1: CdRs format.

	CdR field
All	Phone number
	IMEI
	cell Id
	Timestamp
	Event-type (call/SMS/data)
Call	Call type (MO/MT/IMO/IMT)
	Call duration
	Phone number of the correspondent
SMS	SMS type (MO/MT/IMO/IMT)
	Phone number of the correspondent
Data	Data session size

Table 2: Review of generated CdRs completeness

	Mobility features	Traffic features		
		Call	SMS	data
[32]	✗	✓	✗	✗
[42]	✗	✓	✗	✗
[15]	✗	✓	✗	✗
[29]	✗	✗	✗	✓
[6]	✓	✗	✗	✗
[16]	✓	✗	✗	✗
[49]	✓	✗	✗	✗
[23]	✓	✗	✗	✗
Zen	✓	✓	✓	✓

network cell Ids), all event types, namely data, call, and SMS. Here, the limited access to complete real-world CdRs hardens the modeling and reproduction of complete CdRs. *Zen* circumvents this limitation and provides complete CdRs by jointly modeling separate CdRs features to capture the implicit correlations between them : e.g., the choice of *whom* to communicate with is generally time (*when*) and event (*what*) dependent. Therefore, Table 2 shows *Zen* yields complete CdRs compared to most state-of-the-art contributions which instead provide only one CdRs feature, either mobility or an event type.

Realisticness: *Zen* modeling integrates a real telecom network topology (*inducing users' cell-tower locations*) and organization in multi-operators, as a requirement to produce realistic CdRs. Considering the latter, Zen CdRs conveys inter-operator interaction patterns, valuable for telecom fraud investigation for instance [22].

Fine-grained description: This relates to the realistic reproduction of the individual users behaviors in terms of mobility and traffic, beyond the global behavior of the population. In particular, daily individuals' cellular traffic presents a notable heterogeneity that challenges its reproductions. Fig. 3(left) shows a daily traffic (i.e., sequence of events per user) of 100 randomly selected users from a real-world CdRs. We can see a great diversity of users regarding events generation. Statistical approaches (see Fig. 3(center)), as commonly used in the state-of-the-art [29, 32, 42], are limited in reproducing such traffic dynamics as they do not allow per-user modeling but per-user profile (i.e., group of users with similar behavior). Improving this result, the approach we use in *Zen* better captures such individuals heterogeneity (see Fig. 3 right).

Privacy compliance: *Zen* leverages *refCdRs* with no geographical information associated with traffic events. From such CdRs, *Zen* uniquely captures and reproduces individuals' traffic behavior in time, which can be then associated with any modeling of individuals' daily urban mobility. In *Zen*, mobility behavior is emulated as realistically as possible. Such disjoint modeling hides real individuals' spatiotemporal daily-life habits in routine and leisure times (e.g., home/work, nightlife, etc.), bringing the privacy-preserving capability to the produced CdRs.

3 THE TRAFFIC MODULE

We describe here the generative model used to reproduce CdRs traffic behavior. Our generative model has enough expressive power to capture inter-CdRs feature correlations while considering individual users' behavior. In particular, we leverage an enhanced *recurrent neural network* (RNN), named *Long-Short-Term Memory* (LSTM), known for its ability to generate complex, realistic long-range sequences.

Our model is trained from *RefCdRs* which report a set of timestamped events generated by several users. Each CdRs' event (or a line) includes the following information: start time, user id (i.e., phone number), event-type (i.e., data, SMS, call), corresponding user id (for calls and SMS), call duration (for calls only), and data volume (for data only).

We organize *RefCdRs* by user: the set of events chronologically generated by the user u throughout the trace forms a sequence of events $(e_1^u, e_2^u, e_3^u, ...e_{N_u}^u)$ of size N_u, which is the model basis. Hence, data reproduction is done in a sequential order, i.e., from time step 1 to N_u. The generation of an event in a sequence is a four-stage process, where each stage relies on the previous output.

Stage 1: at step t, we predict the next event-type e_{t+1}^u a user will perform, using the *event-type model* (cf. §3.1).

Stage 2: given the event-type, the *inter-event time (IET) model* generates the IET value used to deduce the starting time for the predicted event-type e_{t+1}^u (cf. §3.2).

Stage 3: the *correspondent model* predicts which of its correspondent a user will interact with for the next event e_{t+1}^u (§3.3). This model is executed only if e_{t+1}^u is a call or SMS, i.e., the only events requiring correspondent interactions.

Stage 4: Finally, the *metric model* refers to how the events are generated: For call events, it generates its duration, while for for data events, it produces the data volume (§3.4). Note that the temporal information is not constant throughout the pipeline. From stages 1 to 2, we use the temporal information of the event-type at step t to predict the one of the event-type at step $t + 1$, then used in stage 3.

3.1 Event-type modeling

The *event-type model* predicts the next event-type a user will generate from four types of events: data, local calls (uniquely outgoing), international calls (outgoing or incoming), and local SMS (uniquely outgoing). Local incoming calls and SMS are modeled here as they are induced from outgoing calls and SMS during the generation. Modeling international calls separately from local calls, rather than having a unique "call" event-type and determining probalistically if it is local or international, allows distinguishing different user behaviors towards international calls. As shown in Fig. 3, some users may not make international calls while others make them

frequently. Finally, we did not model international SMS event-type because it is rare and not present in *RefCdRs*.

The event-type model. We model sequences of event-types using an LSTM. At step t, the LSTM takes as input a vector of features x_t and generates a vector of four scores, $y_t = (y_t^1, y_t^2, y_t^3, y_t^4)$. These scores parameterize a multinomial distribution $Pr(\widehat{e}_t^u|y_t)$ for the next event-type $\widehat{e}_{t+1}^u$, through a softmax function: $Pr(\widehat{e}_t^u|y_t) = \frac{exp(y_t^k)}{\sum_{k'=1}^{4} exp(y_t^{k'})}$.
When training, the true previous event-types at step t are encoded as input for the next step. Network parameters' training is done according to the standard approach of minimizing the negative-log-likelihood of the training data. We compute the gradient of this loss with respect to our network parameters through backpropagation.

Features x_t**.** At step t, we distinguish four features for predicting e_{t+1}^u: the event-type at step t (one-hot encoded) and its temporal features, i.e., Day-of-Week (DOW, one-hot encoded), Hour-of-Day (HOD, one-hot encoded), and Second-of-Day (SOD, cyclical encoded). A one-hot encoding represents the ith of N features using a N-sized vector of all zeros, except for the ith element, which is set to 1. A cyclical encoding maps a continuous inherently-cyclical feature into two dimensions using a sine and cosine transformation. The *HOD* and *DOW* features capture the seasonality and regularity of mobile traffic (less activity at night and during weekends [5]). The fine-grained encoding of time as *SOD* is used to capture the very short temporal difference between consecutive events (e.g., tens of seconds for data events).

3.2 Inter-event time modeling

The *IET model* returns the possible time values between a sequence's events with a confidence interval. It works in two steps: first, we use an LSTM to parameterize a multinomial distribution over a discrete set of time bins. Then, we use statistical methods to sample a continuous value inside a predicted time bin. In the following, we present our considerations for discrete IET estimation, then the detail of our LSTM network, and finally, our methodology for sampling an IET value given an IET bin.

Discrete IET estimation. IET are divided into discrete bins, $b_1, .., b_J$, representing J consecutive intervals of time. To determine the bin boundaries, [24] recommends setting boundaries at evenly-spaced quantiles of time in training data. We found that, in our case, such a setting results in tiny intervals for the smallest values of IET due to the IET's heavy-tailed distribution. For instance, considering the 4-quantiles, there are as many elements in $[1s - 20s[$ as in $[20s - 72s[$. A division at the 20s could distort the model's accuracy while being acceptable for realistic CdRs. Thus, we chose the IET bins empirically to make the model less complex and easier to train without increasing the reconstruction error in mapping back to continuous values. We, therefore, divide IET into three intervals: $[0s - 30min]$ $]30min - 24h]$, and $> 24h$.

The IET LSTM model. The LSTM network takes at each step, t, as input a feature vector, x_t and generates as output a vector of scores y_t, with one score for each possible IET bin. As with the *event-type model*, these scores are used as logits in a softmax to get a multinomial distribution over the time bins. To train the network parameters, we minimize the negative-log-likelihood of the training data.

Table 3: IET distribution and parameters per bin

IET bin	Distribution	Parameters
$[0s - 30min]$	Lognormal	$\sigma = 1.798, \mu = 4.04, x_0 = 0.99$
$]30min - 24h]$	Lognormal	$\sigma = 1.731, \mu = 8.59, x_0 = 1749.08$
$> 24h$	Exponential	$\lambda = 6.21e-06, x_0 = 86401$

Features x_t**.** At each step t, we consider as features, the temporal information of e_t^u (§3.1) as well as the predicted event-type e_{t+1}^u, one-hot encoded.

Continuous estimation. Generating CdRs traffic requires knowing the precise starting time of the next event of the sequence, which is used for further predictions. Therefore, we convert the predicted discretized IET bins to real-values. We apply to each IET bin the KS statistic test to estimate the distribution and related parameters best fitting the corresponding empirical distribution in *RefCdRs*. Table 3 shows the fitted distributions to sample an IET value per bin. The model returns the median value and the confidence interval of the values obtained after n sampling (by default $n = 1$).

3.3 Correspondent modeling

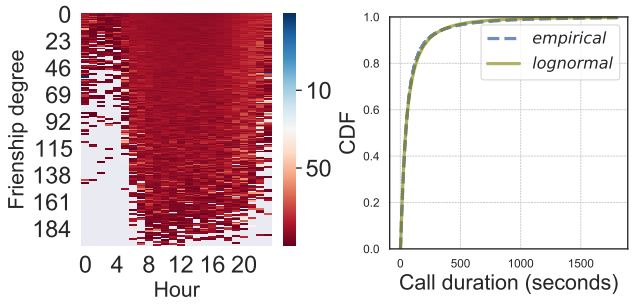
The *correspondent model* applies only for event-types requiring interaction with a correspondent (i.e., SMS and local or international calls). We first define the notion of *friendship degree* (fd), intuitively capturing the friendship strength of a user with each of its correspondents. Let u be a user, with $\#c_u$ correspondents over the considered period, we then call $\#e_c^u$ the number of events the user u had with his correspondent c. We increasingly order the correspondents of u according to their corresponding number of events such that $\#e_1^u \leq \#e_2^u \leq .. \leq \#e_j^u \leq .. \leq \#e_{\#c_u}^u$. The *friendship degree* of the correspondent c of u is the rank j of c in this order. Hence, at step t, the *correspondent model* returns a predicted *friendship degree* $\widehat{fd}_t^u$ for the correspondent with whom the event e_t^u is done.

Correspondent LSTM model. The *correspondent model* is also a LSTM network that takes as input at step t, a feature vector x_t per user. It generates as output the predicted *friendship degree* $\widehat{fd}_t^u$. The network parameters training minimizes the Mean Absolute Error (MAE) of the training data.

Features x_t**.** At step t, the features are: the temporal information of e_t^u (cf. §3.1) except the *SOD*, the one-hot encoded event-type e_t^u, and the number of correspondent of u, $\#c_u$. This later is constant throughout a user sequence and is essential to help the model captures that $\widehat{fd}_t^u \leq \#c_u$. Accordingly, it is not encoded and is left to its actual value.

3.4 Metric modeling

This section presents the models used to generate the metrics (i.e., a model per metric) associated with events generation, namely the call duration and the data volume.



(a) (b)

Figure 4: (a) Avg call duration (s) per hour and friendship degree (b) Call duration CDF for *RefCdRs*.

Call duration We use a statistical method to model the call duration. In fact, contrary to the previously modeled parameters, there is no explicit features dependency or variability (and therefore, no complexity) regarding call durations, which implies that a used RNN could hardly train. This is confirmed in Fig. 4a, which shows the variation of the average call duration per hour and per friendship degree over the entire dataset. We can see that overall, call duration does not vary much, and thus, there is no particular correlation between these parameters. Moreover, the per-user behavior regarding call duration (easily assessed through the average call duration per user) closely depends on the number of calls each user makes over the CdRs duration, which is opportunely already captured by the *IET model*. Accordingly, the *call duration model* corresponds to the estimation of the parameters of the continuous distribution that best fits the empirical distribution of call duration, as shown in Fig. 4b. From a statistical test, we found this distribution to be Lognormal of parameters $\sigma = 1.29, \mu = 3.78, x_0 = -0.47$.

Data volume. The *data volume model* returns a data volume value for each data event. According to 3GPP standards, each data-typed CdRs line corresponds to the generation of a data session by a user. Unfortunately, as *RefCdRs* lack this information, we rely on the study done in [29] to design the *data volume model*. To the best of our knowledge, [29] is the only work that conducts a thorough characterization of data volume usage per session and per user over time extracted from real-world CdRs, as well as designs a generator of realistic CdRs that conforms to these characterizations.
[29] profiled users' data usage over time according to their generated amount of data (*volume profile*, i.e., Light, Medium, or Heavy) and to how often they generate data sessions (*frequency profile*, i.e., Occasional or Frequent). Besides, it extracted from real-world CdRs the distributions of data session volume according to a user's profile and the day period (peak or off-peak hours) and the percentage of users per profile. We use such percentages to first assign a *volume profile* to each user in *Zen*. As the *frequency profile* could be inconsistent with the frequency of data event-type as predicted by the *event-type model*, we attribute to each user, in *Zen* the *Occasional frequency profile*. In fact, the distribution of the number of data sessions per day and user from *RefCdRs* shows the majority of the population to be of this latter profile. Finally, we sample from the distributions found in [29] to get a data session volume.

4 THE MOBILITY MODULE

The *Zen's mobility module* produces realistic CdRs mobility traces in three steps each covered by a sub-module. The *mobility-generator* (§4.1) emulates a population urban mobility in a real-world city map, with users displacements generated according to public sources [34] and describing city planning and transportation information [17]. Next, from the input city map, the *topology-builder* (§4.2) builds a realistic cellular topology using cell towers' positioning of mobile operators deployed in the considered real-world city, gotten from OpenCellID [33]. This topology is then used in the *position-to-cellId module* (§4.3) to map the mobility traces produced by the *mobility-generator* to the cell granularity.

4.1 Mobility generator

Zen mobility-generator inherits the highly configurable capability of the *Opportunistic Network Environment* (ONE) [20] simulator. Besides, it enhances the *Working Day Mobility* model (WDM) [9] of ONE into a model named *En-WDM*, and generates CdRs of format `<Timestamp, userId, lat,lon>`.

Our motivation to use WDM as a basic building block of *Zen* is twofold. First, contrary to related models [41, 44], WDM originality comes from the combination of various mobility aspects present in people daily life (e.g., home and workplaces, day periods). Second, WDM closely reproduces wireless interactions (i.e., inter-contact and contact time) distributions found in two real-world measurement experiments (i.e., iMote and Dartmouth), asserting modeling generality.
Nevertheless, WDM is limited in capturing some fine-grained real mobility habits or fine-tuning. *En-WDM* tackles such limitations and strengthens the model with additional literature's intuitions on laws dictating human mobility behavior, such as preferential attachment, regular daily behavior, transportation-dependent shortest-path preferences, and most importantly, uncertainty (i.e., novelty-seeking behaviors) and heterogeneity. Next, we detail *En-WDM*.

WDM's inherited functionalities. *En-WDM* models week working days' movements into three activities and their transitions, i.e., "home", "working", and "night activity". The night activity corresponds to leisure-related times spent in preferred spots of friends groups.

Exploration profiling. Users in *En-WDM* emulation decide in a probabilistic-way whether to go home or to a night activity. To setup such probabilities, we rely on the exploration phenomenon profiling conducted in [1] and define three mobility profiles: *scouters* are more inclined to explore and discover new places to visit, *routiners* rarely explore and prefer to stay among their familiar and few known places, and *regulars* constantly alternate between exploration and routine. We then accordingly classify users given by the *ChineseDB* dataset (cf. Sec 2.3) in these three profiles. Results describe a population with 20.27% of *scouters*, 54.75% of *regulars*, and 24.98% of *routiners*. After this classification, we assign to users in each profile, the probabilities of "nightlife activity": 0.8 for *scouters*, 0.5 for *regulars*, and 0.2 for *routiners*.

Neighborhood and popularity. Rather than considering home/office locations' (lat, lon) coordinates, *En-WDM* associates each location coordinate to the center of a neighborhood of rectangular shape

Table 4: Key parameters for *En-WDM* emulation

***En-WDM* Parameter Description**	**Value vs Default**
Size of the office squared-shaped side	*100*
Minimum size of a friends group for evening activities	*1*
Maximum size of a friends group for evening activities	5 vs 3
Minimum value for evening activities duration	1h vs 10s
Maximum value for evening activities duration	4h vs 2h
Probability for a user to own a car	0.19 vs 0.5
(width, heigth) of the emulation area	*(10000, 8000)*
(width, heigth) of a home cluster	**(250, 150)**
(width, heigth) of a office cluster	**(500, 300)**

and configurable size. This allows to distinguish areas with high housing density (e.g., residential areas, university campuses), areas with high business density (business districts), and popular leisure locations. A user is first assigned a home/office neighborhood and then, chooses her exact home/office location randomly inside the neighborhood. Moreover, we added the *neighborhood popularity*, which represents the probability for a user to choose a given neighborhood as a home/office neighborhood or, in the case of night activity, the probability of choosing a spot for her evening activity.

Distance-based profiling. *En-WDM* enables the definition of cities' districts (hereafter, areas) to replicate the real world. Accordingly, we associate each user to one of the three profiles representing area displacements: *profile 1* inside a single area, *profile 2* among two areas, and *profile 3* in the whole map. To get the population percentage to be considered in each profile, we profile *Geolife*'s users resulting in: *Profile 1* including 72% of users whose maximum distance is less than 1/3 of the maximum observed distance D_{max} ($\approx 2.49 \times 10^3 km$). *Profile 2* with 19% of users with a maximum distance between 1/3 and 2/3 of D_{max}, and *profile 3* including 9% of users with a maximum distance greater than 2/3 of D_{max}.

Simple parameterization. We report here all the key configuration parameters needed for *En-WDM* emulation. Table 4 summarizes them. We use italic style for those we used the default value and regular one for those we modified. We set the value of *ProbOwnCar* to 0.19 based on transportation statistics in the city of Helsinki [17]. Parameters in bold (*homeRange* and *officeRange*) are those we added for clusters implementation. In particular, the ratio between the *worldSize*, *officeSize*, and cluster sizes may vary depending on the emulated city. These parameters values in Table 4 are adapted for a emulation in the city of Helsinki.

4.2 Topology builder

The *topology-builder* uses the geographical positions of base stations (BS) in the emulated area, as given by OpenCellId [33], and performs a Voronoi tesselation. The tessellation produces a cellular network topology with heterogeneous cell sizes close to reality, containing each input BS. Each Voronoi cell defines the communication boundaries of an input BS. For generality and simplicity reasons, we include all operators' base stations given by OpenCellId in a bigger architecture to derive the Voronoi topology. This unique topology is assigned to all operators considered in *Zen*'s process. In practice, sharing BSs between different operators is commonly done for cost savings.

4.3 Position-to-cellId module

The *position-to-cellId module* assembles the modeled users' mobility and the designed Voronoi cellular topology. For this, each user's geographical position given by the *mobility-generator* traces is mapped to the corresponding OpenCellId's BS identifier, i.e., cellID, in the Voronoi topology. It outputs mobility CdRs in the format `<Timestamp, userID, cellID>` describing users' spatiotemporal daily mobility in a real city map and adapted to a real network topology. Despite the realism given by such leveraged real-world information, the generation of users' mobility has a realistic and not a real nature since no ground-truth information on users' real-life routine is available. This brings privacy benefits to *Zen* CdRs.

5 THE *SOCIAL-TIES* MODULE

Zen CdRs generation lays on the *social-ties module* providing the network social structure. This structure induces phone numbers from users of the mobility CdRs and builds the network social graph by creating per user's phonebook, i.e., the users she can interacts through calls or SMS.

Mobility users to phone numbers. From the number of network's operators and the users distribution per operator (taken as parameter or induced from OpenCellId[33]), the *social-ties module* assigns an operator per user and generates a phone number in the format `<MCC><MNC> <5 random digits>`, where MCC and MNC describe the mobile code for country and the operator network code within the country, respectively.

Network social graph. Reproducing the social graph of users' interactions implies answering the following three questions. The term "correspondent" refers to a phone number in a user's phonebook.

(Q1) how many correspondents does each user have? To answer this question, the *social-ties* module relies on the distribution of correspondents per user from *RefCdRs*. Let $u \in U$ be a user with $\#c_u$ correspondents; we consider the non-parametric distribution $P_{\#c} = P(\#c_u = \#c) \quad \forall\ \#c \in [1, MAX]$. Thus, for each generated user u', its number of correspondents $\#c_{u'}$ is obtained with the multinomial distribution of parameters $P_{\#c}$.
We then define four disjoint categories of correspondents: international correspondents (c_{inter}), outgoing local correspondents (c_{out}), incoming local correspondents (c_{in}), and both outgoing and incoming local correspondents (c_{both}). Thus, $\forall\ u \in U,\ \#c_u = \#c_{inter,u} + \#c_{out,u} + \#c_{in,u} + \#c_{both,u} = (x_{inter,u} + x_{out,u} + x_{in,u} + x_{both,u}) \times \#c_u$. We export the average values $\overline{x_{cat,u}} \quad \forall\ cat \in \{inter,\ out,\ in,\ both\}$. Then, we use the multinomial distribution of $P = \overline{x_{cat,u}}$ to induce the number of correspondents, in each category, of each user.

(Q2) how do we choose these correspondents? We create user phonebooks by implementing a variant of the configuration model algorithm [45], which allows building a graph from given users degrees. We apply this algorithm by correspondents' category so that each user is an c_{in} correspondent of its c_{out} correspondents and a c_{both} of its c_{both} correspondents. Moreover, we add a heuristic to choose users' correspondents based on their relationship type, (i.e., neighbors, colleagues, or friends) extracted from the generated mobility dataset (cf. §4) as follows. users located inside the same home/work cluster between *1am to 4 am* and *10 am to 2 pm*, over the whole dataset duration, are considered neighbors and colleagues,

respectively. As well, users in the same group for night activities, when they occur, are considered friends. Hence, a user's correspondents are selected according to defined probabilities (taken as parameters) from its list of neighbors, colleagues, friends, and other users until we reach the fixed number of the user's correspondents.

At last, the *social-ties* module outputs each user's list of correspondents organized in the categories c_{out}, c_{both}, and c_{inter}, while c_{in} category is induced from the c_{out} one.

(Q3) how does a user interact with all of its correspondents? While (Q1) and (Q2) are tackled by the *social-ties module*, question (Q3) is addressed through the *correspondent model* of the *traffic module* detailed in section 3.3.

6 THE *CDR-COMBINER* MODULE

Zen's *CdR-combiner* module integrates outputs of previous modules to produce realistic CdRs, as follows.

Using *event-type* and *IET models* from the *traffic module*, the *CdR-combiner* generates timestamped sequences of events over the total duration. Then, each sequence is associated with a correspondent determined by the *social-ties* module, based on each user's number of correspondents per category, indicating which event-types the user can generate. At this point, using the *correspondent model*, the *CdR-combiner* predicts a correspondent friendship degree per user event that is later associated with the corresponding phone number from users' phonebooks.

Next, we add complementary metrics to users' events. For all calls events, the call duration metric relates only to available correspondents of users. We do not consider unavailable users' correspondents (i.e., already in an ongoing communication) at the caller-callee association. Hence, for available correspondents, a call duration value is sampled from the *call duration* model distribution. This value is upper-bounded by the time to the closest scheduled call. As well, for data events, the data volume metric is assigned according to the *data volume* model.

Following, the *CdR-combiner* integrates CdRs spatial information, i.e., corresponding users' cell Ids at each event timestamp (resulting from the *mobility module*). At last, based on users' phone numbers, the *CdR-combiner* infers CdRs traces produced by each operator. *Zen*, therefore, generates a complete and realistic CdRs trace per generated mobile operator in the format specified in Table 1.

7 EVALUATIONS

This section confirms *Zen*'s validity by evaluating traffic and mobility modeling separately, then their merging into CdRs.

7.1 Traffic module

Hereafter, we evaluate the accuracy and the performance of predictions resulting from the *traffic module*'s stages. As there is no similar contribution in the literature, we compare *Zen*'s models to designed baseline predictors. Table 5 summarizes all comparison metrics and provides their distributions on the right of each evaluation result.

7.1.1 ***Experimental datasets.*** We train and evaluate our models on *RefCdRs* after some data handling. First, we only consider events of *users subscribed* to the operator network collecting *RefCdRs*. Then, we filter out users having less than 3 generated events in the whole period of 4 weeks and those with more than one event at the same timestamp. Those manipulations result in the selection of nearly 6000 users totalizing 1,782,829 events or CdRs entries, i.e., 77.8% of the *RefCdRs*' initial size. We then use as *training set* the first two weeks of the dataset, the 3rd week as *validation set*, and the 4th week as the *test set*. Because our traffic predictions are user-based, the non-filtered remaining users compose all the three previous sets and only their event sequences varies according to the week considered in each set.

7.1.2 ***Models training and Hyper-parameters.*** We used a 2-layer LSTM with 50 hidden units per layer for the *event-type model* and 100 hidden units per layer for the two other models. To avoid over-fitting the training dataset, we used a dropout regularization with $p = 0.2$. The LSTM losses are iteratively minimized using mini-batch gradient descent with the Adam optimizer. Each mini-batch contains 64 sequences of events (i.e., users). We chose event sequences' lengths of 302 for training, 157 for validation, 159 for test, sampled from the distribution of the number of events generated by users in each experimental set. Therefore, we pad all sequences to the sequence length in each experimental set to homogenize datasets and ease the training. We use a masking layer to tag added values in each sequence to ignore them in the loss calculation. Besides, we fixed a gradient clip value of 0.01 to avoid "exploding gradients" prone to affect RNN.

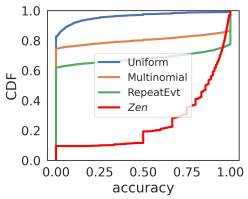
7.1.3 ***Event-type model.*** We compare our *event-type model*'s predictions (cf. §3.1) to the ones of the following baselines: *Uniform* – each event-type is equally likely to occur at each time step; *Multinomial* – each event-type probability is given by its empirical count in training data; *RepeatEvt* – the next event-type is always predicted to be the same as the previous one. We use the following evaluation metrics: (*NLL*) Negative-log-likelihood of next-step probabilities, and (*Accuracy*) next-step 1-best correct classification rate (for this metric, the traditional Multinomial approach always output the most frequent event-type). Results are presented in Table 5. Selecting event-type according to the *Multinomial* is significantly more predictive than the *Uniform*, but worse than *RepeatEvt*. Our *Zen*'s *event-type model* works the best. For both NLL and Accuracy, *Zen* is significantly better than *RepeatEvt*, i.e., the most probable event-type is not always the previous one.

7.1.4 ***IET model.*** As before, we compare the acuteness of our model in predicting the next IET Bin (cf. §3.2) with the corresponding above-defined baselines. Table 5 shows that for both metrics, NLL and Accuracy, the performance of *Zen*'s *IET model* is much higher than *RepeatBin* (that simply repeats the previous IET Bin), followed by the Uniform and the *Multinomial* baselines. Disregarding the prediction approach, we compute the discretized probabilities of IET Bins and map them to IET values in a continuous domain: named *Bin sampling* mapping. To evaluate how efficient *Zen*'s and baselines' *Bin sampling* are, we compare them to the *Overall sampling* mapping, both described next.

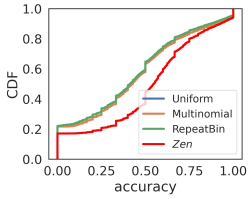
- *Bin sampling*: At each Bin, the IET value is obtained after averaging $n = 500$ samplings of the corresponding continuous IET distribution (see §3.2). We apply this approach to all the previously

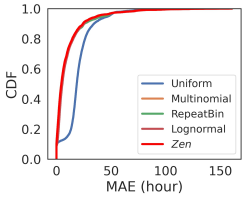
Table 5: *Traffic LSTM models evaluation results.*

Event-type model		
Predictor type	NLL	Accur.
Uniform	0.27	2.91%
Multinomial	0.21	38.97
RepeatEvt	N/A	43.27
Zen	**0.037**	**91.82**

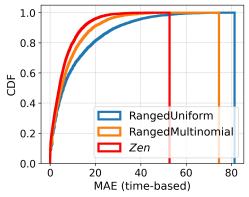


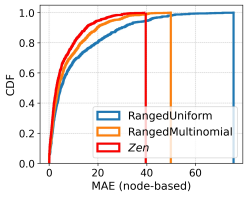
IET model						
Predictor type	NLL	Accur.	MAE	MAE]0, 30 min] (82.8%)	MAE]30min, 24h] (15.45%)	MAE >24h (1.75%)
Uniform	0.215	64.56	1097	1033	1120	2319
Multinomial	0.165	64.56	231	68	334	2877
RepeatBin	N/A	58.05	239	73	347	2871
Lognormal	N/A	N/A	249	78	361	2973
Zen	**0.118**	**69.25**	**185**	**16**	**295**	**2904**





Correspondent model								
Predictor type	MAE (time-based)				MAE (user-based)			
	All	[1,6]	]6,21]	>21	All	[1,6] (50%)	]6,21] (30%)	>21 (20%)
Ranged-Uniform	25.12	1.17	5.04	29.57	26.38	1.08	4.53	31.17
Ranged-Multinomial	15.78	0.91	3.87	18.42	17.28	0.81	3.41	20.38
Zen	11.81	0.65	3.02	13.77	13.23	0.63	2.57	15.68





Bin-based models, i.e., *Zen*'s *IET model*, Uniform, Multinomial, and RepeatBin predictors.

- *Overall sampling*: We perform a fitting of the empirical IET distribution (i.e., with no bins) and obtain a Lognormal distribution with $\sigma = 2.67, \mu = 4.97, x_0 = 1$. Then, we straightly predict continuous values by sampling the resulted fitted IET distribution. We name this prediction *Lognormal*.

The Mean Absolute Error (MAE) of the IETs in minutes is used as the comparison metric. It estimates the average distance between actual and predicted IET. From Table 5, we can notice that the *Bin-sampling* of *Multinomial* and *RepeatBin* have comparable MAE performances, followed by the *Overall-sampling Lognormal* predictor. This behavior is also verified per Bin (three last columns). Overall, *Zen* works the best. In the first bin $]0, 30min]$, which is the most sensitive, we note that except for the *Zen*, all models on average predict an IET value outside the initial interval.

7.1.5 ***Correspondent model.*** At last, we evaluate the *correspondent model* (cf. §3.3) by comparing its predictions to the following baselines:

- *RangedUniform*: Per user u, correspondents c_i, $\forall i = 1, 2, ..., \#c_u$ are equally likely to be predicted at each sequence step.
- *RangedMultinomial*: Per user u, each correspondent c_i is chosen with a probability $(p_i^u, 1 \leq i \leq \#c_u)$ extracted from the procedure as follows:
  Let U be the set of users and u a user in U. We recall that $\#e_{c_i}^u$ refers to the number of events u made with his correspondent c_i. From this definition, we derive $P_{c_i}^u$ the proportion of events made by u with its correspondent $c_i : P_{c_i}^u = \#e_{c_i}^u / \sum_i \#e_{c_i}^u$.
  For all $i = 1, 2, ..., MAX(\#c_u)$ we extract the mean values $\overline{P_{c_i}} = \overline{P_{c_i}^u}\ \forall u \in U$. Hence, for a user u, the probabilities $(p_i^u, i = 1, 2, ..\#c_u)$ is obtained by normalizing the first $\#c_u$ mean values $(\overline{P_{c_i}}, i = 1, 2, ..\#c_u)$ such that $\sum_i p_i^u = 1$.

The evaluation metric is the MAE of the predictions $\widehat{fd}_t$ in the test dataset. We found that as we train the *correspondent model* with chronologically-separated experimental windows (defined in §7.1.1), the MAE loss value continually increases in the validation dataset. This is due to the fact that in the training period (i.e., first two weeks), users only interact with some of their correspondents, making it difficult for the model to generalize. To fix this issue, we instead split training, validation, and test datasets by selecting users traffic over the whole dataset period (4 weeks). The training dataset includes 60% of the users, while the validation and test datasets each represent 20%. Results in Table 5 show the *Ranged-Multinomial* predictor has significantly better results compared to the *RangedUniform* predictor.

Overall, *Zen* is the modeling that best performs, showing its ability to capture users interaction with their correspondents. In particular, the detailed distribution plots show *Zen* presents for 80% of users (i) more than 95% and 75% of accuracy for respectively, the event-type and IET models, and (ii) less than 6.68% and 12.5% of MAE maximum values for respectively, the IET and correspondent models.

7.2 Mobility module

We validate our *En-WDM* mobility model by comparing it to its original version, the WDM [9]. We rely on WDM results closely following real-world measurement datasets distributions (i.e., iMote or Dartmouth). Since *En-WDM* adds new functionalities in modeling mobility to WDM, we are not looking for identical results from both models but for similarities in terms of distributions and curve behaviors.

Fig. 5 shows well-known metrics for characterization of wireless networking meetings (inter-contact and contact time) and the tendencies in human mobility, i.e., confinement (radius of gyration) and repetitiveness (probability to return to previously visited places). As for WDM, we emulate a scenario with 1000 and 6000 users, moving in the Helsinki city center area with roughly 7×8.5km for $5.10^5 s$ and with the same arrangement of home/work and POIs. We use the same representation of results for comparison reasons.

We can see that *En-WDM*'s inter-contact time distribution (cf. Fig. 5a) and the normalized number of contacts (Fig. 5b) closely follows the ones of WDM, attesting the realistic modeling of such metric at population scale and the capability of reproducing heterogeneity to mobility decisions.

At last, we evaluate the capability of the two models in reproducing seminal literature analytical human mobility laws [1, 11, 30]. The radius of gyration (Fig. 5c) estimates the area size mostly covered by daily displacements of a user. In *En-WDM*, the radius of gyration is globally smaller due to routiners and regulars (79.73% of the population) who have more confined displacements, consistent with real-life mobility behavior [1]. Moreover, the average return probability (Fig. 5d) and per-cell repetitiveness ((Fig. 5e) results show that users have a regular and periodical spatial mobility behavior with a higher probability of returning to a previous small set of visited locations, as shown in [11, 30].

7.3 *Zen* CdRs use cases

We evaluate the complete CdRs resulting from *Zen* framework as compared to *RefCdRs* when applied to three use cases. As *RefCdRs* lack ground-truth in mobility information, we enrich them with *Zen* CdRs' emulated user trajectories using *Zen*'s CdR-combiner methodology (ref. §6), we name it *M-RefCdRs*. Based on the confirmed *Zen* performance in reproducing human mobility laws, we focus our use-cases analysis on the reproduction of cellular traffic behavior for which we have a ground-truth. We generate *Zen* CdRs with 6000 users, corresponding to the same number of users in *RefCdRs* (see §7.1.1) and consider a week-long period.

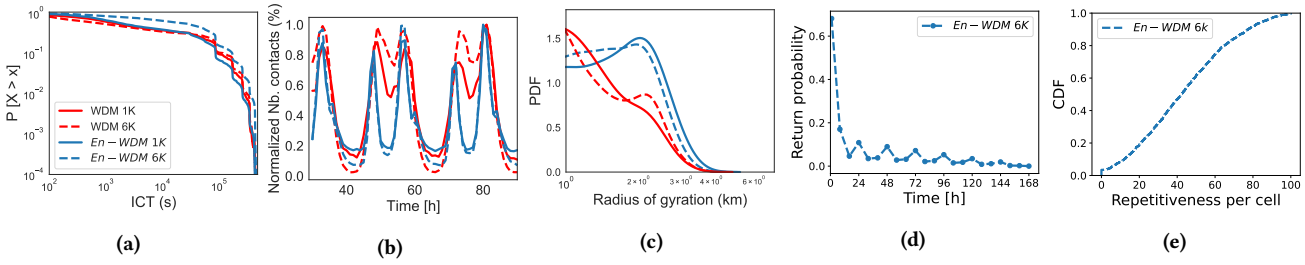
Dynamic urban tracking. Real-time population density tracking is a key functionality to support adaptive urban and transport planning. As shown in [21], such density at time t can be derived from the corresponding network activity load at t computed as the mean number of network events (e.g., here ongoing calls, exchanged SMS, and established data sessions) per individual. Following this methodology, Fig. 6 shows the spatial distribution (values in the color bar) of people presence in network cells of an Helsinki area (2.2km × 3.6km), at four representative time hours of individuals' routine, obtained with *M-RefCdRs* and *Zen* CdRs. As in *M-RefCdRs*, we see that people presence at the office period (8h-12h) is concentrated in specific zones corresponding to defined Helsinki business neighborhoods. In contrast, the after-work period (18h-22h) includes displacements times and night activities not made at specific spots (e.g., groups of users can walk down the streets for their night activity), explaining people presence is spread over a broader zone. Besides, we notice that people presence is captured equivalently in *M-RefCdRs* and *Zen*'s CdRs, especially in working period (8h-12h). We believe the resulting few dissimilarities, particularly for the after-work period (18h-22h), are mainly due to the non-deterministic association of user's traffic to trajectories in Zen's CdR-combiner.

Data-Driven Micro BS Sleeping. Numerous works studied power savings in Radio Access Networks (RAN). Inspired by [47], we investigate how a traffic-aware Base Station (BS) on/off-switching strategy [43] performs when informed with *Zen* CdRs compared to *M-RefCdRs*. We assume an heterogeneous RAN deployment where each cell is served by a separate micro BS, whereas macro BSs provide umbrella coverage to a larger area. Specifically, we consider a grid tessellation of 5X5 macro BSs in the considered zone. The power needed to the operation of a BS at time t is $P(t) = N_{trx}(P_0 + \Delta_p P_{max}\rho(t))$, $0 \leq \rho(t) \leq 1$, where $\rho(t)$ is the relative traffic load at time t with P_0, N_{trx}, P_{max} and Δ_p being constants defined for micro and macro BSs in [47]. Then, if $\rho(t) \leq \rho_{min} = 0.37$ as considered in [8] the micro BS offloads its local traffic to the macro BS and goes into sleep mode, where it consumes negligible power. Accordingly, Fig. 7 shows the power consumption ($P(t)$ values in the color bar) of each cell's micro BS at two hours in Helsinki (a zoomed-in area of 2.2km × 1.6km) with and without such a strategy implemented. We can see that comparable cells are kept on, while the strategy brings similar energy savings.

Anomaly detection. Beyond global population-related applications, the fine-grained state of *Zen* CdRs allows for the investigation of per-user spatiotemporal behavior for cellular anomaly detection. Such anomalies can be unusual events possibly generated by some security incidents (e.g., stolen account, malware device infection) [10] or users with a fraudulent behavior profile. As an instance of the latter, SIMBox fraud is a prevalent scam in telecommunication networks consisting of "fake" user accounts re-injecting diverted international calls as local calls to a country [22]. We assess the utility of *Zen* CdRs for investigating such fraud by applying a user profiling method where traffic or mobility users' behaviors are leveraged to classify a user as fraudulent or not. To this end, we apply for both *Zen* and real ones, a DBSCAN clustering to a set of per-user traffic-related features specific to detect SIMBox fraudulent behavior as described in Table 1 of [40]. Results show a similarity between *Zen* CdRs and real-world ones: while *M-RefCdRs*' estimated number of clusters and outliers are 10 and 1241, Zen CdRs' confidence intervals for these metrics are 9.1 ± 1.66 and 1122.3 ± 35.02 for 10 samples of *Zen* CdRs' call duration feature (ref. §3.4).

8 RELATED LITERATURE

CdRs' inaccessibility has pushed researchers to generate their own synthetically, commonly using features modeling. This leads to either mobility- or traffic-specific CdRs, often with grouped-based analysis of individuals' behavior. *Zen* tackles such lacks by empowering the scientific community with the autonomy needed for the generation of realistic, complete, precise, and flexible CdRs.



(a) (b) (c) (d) (e)

Figure 5: Mobility metrics compared with initial WDM: (a) Inter-contact time CCDF (b) Normalized number of contacts per hour (c) Radius of gyration CDF (d) Return probability (e) Per-cell repetitiveness CDF

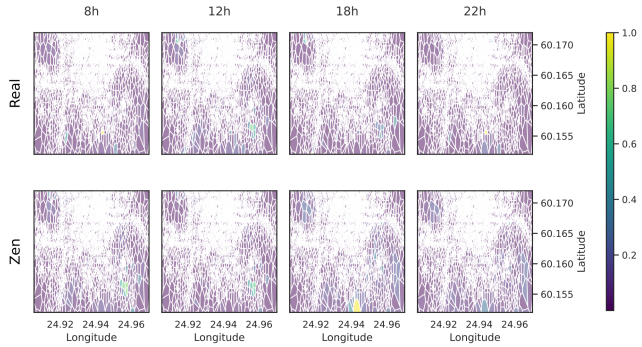
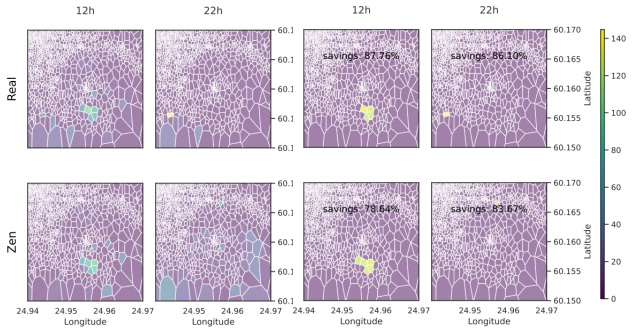


Figure 6: Dynamic people presence estimated at four daily time in Helsinki for real-world and *Zen* traffic.



(a) Always-active micro BS. **(b) Cell-sleeping strategy.**

Figure 7: Power consumption per cell (a) for always-active micro BS and (b) with a cell-sleeping strategy.

Traffic-related: Instead of aggregated network traffic generation as done in [25, 46, 47], we focus here on per-individual CdRs generation that tackles different challenges. In this domain, literature's synthetic CdRs lack completeness in describing both call [15, 32, 42] and mobile data [29] usages, and to the best of our knowledge, pay no attention to SMS usage. Murtić et al. [32] used Social Network Analysis to reproduce call behaviors' features (i.e., temporal likelihood of calls and call duration distribution) per user profile, extracted from real-world CdRs. Nevertheless, the work did not include any validation. In the same vein, Songailaitė et al. [42] statistically model key parameters from real CdRs to produce realistic CdRs. Calling behaviors is simulated based on the empirical fitting of call duration, call count, call likelihood per hour, and weekdays similarity in behaviors. However, simulation relies on a simplistic and randomly-built network social structure leveraging static parameters such as the maximum number of friends and acquaintances. Using a GAN generative model, Hughes et al. [15] show the deep learning models' capability to learn inherent and complex distributions from real CdRs. Unfortunately, real and generated CdRs included only two features: the starting call hour and duration in minutes, revealing a limited extent of modeled features compared to complete CdRs. Finally, Oliveira et al. [29] focused on the data-traffic profiling, modeling, and generation from real CdRs. Their model allows generating data usage's timestamped records per profiled user. Although providing flexible settings for profiles' granularity, this work also has the drawback of modeling only data traffic features, lacking thus real-world CdRs' completeness.

Mobility-related: Synthetically generated mobility traces are regular in literature and frequently extracted from models implemented in ONE [20], BonnMotion [2], or SUMO [26] realistic simulators. Several works on mobility modeling actually focus on the generation of synthetic traces that capture specific features in human mobility that are often domain-specific: e.g., MANETS and DTNs (e.g., inter-contact and contact time) [31, 41], Disaster Management [3, 36] or Sociology [4]. Still, a few works such as [9, 12, 19, 37] aim to model real-life mobility and propose more complex models, valuable for more applications. This paper leverages the [9]'s originality in combining various mobility aspects and realistically modeling them. Other strategies rely on recurrent neural networks [23] or statistical generative models based on real mobility traces such as Markov models [6], spatiotemporal empirical distributions [16] or travel demand [49]. Yet, only a few works [13, 27] address the privacy issues of generated mobility traces, which is however crucial.

9 CONCLUSION AND DISCUSSION

This paper presented *Zen*, the first framework allowing the autonomous generation of complete and realistic CdRs in an individual basis. To this end, we relied on a fully anonymized and incomplete (only traffic-related) CdRs datasets and provide the first literature modeling that captures long-range and inter-CdRs traffic features

correlation, individuals heterogeneity and social-ties in communication. The disjoint modeling of realistic emulated mobility and captured real-world traffic behaviors hides real individuals' daily-life habits in routine and leisure times (e.g., home/work, nightlife, etc.), bringing the privacy-preserving capability to the produced *Zen* CdRs. Finally, we validate *Zen* Cdrs (i) realisticness in reproducing daily cellular behaviors of urban population and (ii) usefulness in practical networking applications such as dynamic population tracing, Radio Access Network's power savings, and anomaly detection as compared to real-world Cdrs. Next, we provide extra discussions on possible alternatives and improvements.

Flexibility and generalization: All the contextual building blocks feeding the *Zen* mobility modeling (e.g., Census information, bus schedule, real city map, neighborhood popularity, etc.) bring generality and flexibility to the representation of city urban life, yielding individuals' cyclic behavior. On the other hand, though *Zen* provides realistic traffic behavior models trained from a unique real-world traffic CdRs, the modeling methodology of this paper is general and can be applied to other CdRs with different cultural traffic habits.

Alternative modeling approaches: LSTMs are perhaps the simplest network (in terms of manual tuning) that can reliably model long-term dependencies and has the flexibility to be used jointly with other more complex architectures. For example, a GAN [14, 47] uses paired generator/discriminator networks to enable very realistic output; our work provides the networks that can be used inside the GAN.

Future improvements: As mentioned, *Zen* extensively enhances the original WDM model. Nevertheless, as with any research contribution, the mobility generation of *Zen* is still open for improvements, such as the addition of complementary features in fine-grained human mobility laws, cities' contextual information (e.g., friendship, popular leisure zones in the city, etc.), representing weekend mobility and behaviors induced by teleworking or minor users profiles (e.g., unemployed), or modeling from real-world mobility CdRs.

Privacy vs individual precision: Although presenting ground-truth modeling and validation opportunities, individual-based mobility modeling of real-world CdRs brings important privacy issues: As users' actual habits in mobility are captured in the model, the generated CdRs have the weakness of revealing aspects in users' daily-life routine, such as important locations (e.g., home/work), regular trajectories (e.g., preferred places for leisure, etc). The tradeoff between privacy compliance and preciseness in mobility modeling of CdRs is a relevant investigation we let for future work.

REFERENCES

[1] Licia Amichi, Aline Carneiro Viana, Mark Crovella, and Antonio A.F. Loureiro. 2020. Understanding Individuals' Proclivity for Novelty Seeking *(SIGSPATIAL '20)*. Association for Computing Machinery, New York, NY, USA, 314–324. https://doi.org/10.1145/3397536.3422248

[2] Nils Aschenbruck, Raphael Ernst, Elmar Gerhards-Padilla, and Matthias Schwamborn. 2010. BonnMotion: A Mobility Scenario Generation and Analysis Tool. In *Proceedings of the 3rd International ICST Conference on Simulation Tools and Techniques* (Torremolinos, Malaga, Spain) *(SIMUTools '10)*. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), Brussels, BEL, Article 51, 10 pages. https://doi.org/10.4108/ICST.SIMUTOOLS2010.8684

[3] Nils Aschenbruck, Elmar Gerhards-Padilla, Michael Gerharz, Matthias Frank, and Peter Martini. 2007. Modelling Mobility in Disaster Area Scenarios. In *Proceedings of the 10th ACM Symposium on Modeling, Analysis, and Simulation of Wireless and Mobile Systems* (Chania, Crete Island, Greece) *(MSWiM '07)*. Association for Computing Machinery, New York, NY, USA, 4–12. https://doi.org/10.1145/1298126.1298131

[4] V. Borrel, F. Legendre, M. Dias de Amorim, and S. Fdida. 2009. SIMPS: using sociology for personal mobility. *IEEE/ACM Transactions on Networking* 17, 03 (may 2009), 831–842. https://doi.org/10.1109/TNET.2008.2003337

[5] Julián Candia, Marta C González, Pu Wang, Timothy Schoenharl, Greg Madey, and Albert-László Barabási. 2008. Uncovering individual and collective human dynamics from mobile phone records. *Journal of Physics A: Mathematical and Theoretical* 41, 22 (May 2008), 224015. https://doi.org/10.1088/1751-8113/41/22/224015

[6] Juan Gonzalo Cárcamo, Roderick Grahm Vogel, Adam M. Terwilliger, Jonathan Leidig, and Greg Wolffe. 2017. Generative models for synthetic populations. In *SummerSim*.

[7] Hsiao-Han Chang, Meng-Chun Chang, Mathew Kiang, Ayesha Mahmud, Nattwut Ekapirat, Kenth Engø-Monsen, Prayuth Sudathip, Caroline Buckee, and Richard Maude. 2021. Low parasite connectivity among three malaria hotspots in Thailand. *Scientific Reports* 11 (12 2021). https://doi.org/10.1038/s41598-021-02746-6

[8] Mattia Dalmasso, Michela Meo, and Daniela Renga. 2016. Radio Resource Management for Improving Energy Self-Sufficiency of Green Mobile Networks. *SIGMETRICS Perform. Eval. Rev.* 44, 2 (sep 2016), 82–87. https://doi.org/10.1145/3003977.3004001

[9] Frans Ekman, Ari Keränen, Jouni Karvo, and Jörg Ott. 2008. Working Day Movement Model. In *Proceedings of the 1st ACM SIGMOBILE Workshop on Mobility Models* (Hong Kong, Hong Kong, China) *(MobilityModels '08)*. Association for Computing Machinery, New York, NY, USA, 33–40. https://doi.org/10.1145/1374688.1374695

[10] Paul Giura, Ilona Murynets, Roger Piqueras Jover, and Yevgeniy Vahlis. 2014. Is It Really You? User Identification via Adaptive Behavior Fingerprinting. In *Proceedings of the 4th ACM Conference on Data and Application Security and Privacy* (San Antonio, Texas, USA) *(CODASPY '14)*. Association for Computing Machinery, New York, NY, USA, 333–344. https://doi.org/10.1145/2557547.2557554

[11] Marta C. González, César A. Hidalgo, and Albert-László Barabási. 2008. Understanding individual human mobility patterns. *Nature* 453, 7196 (jun 2008), 779–782. https://doi.org/10.1038/nature06958

[12] Michal Gorawski and Krzysztof Grochla. 2013. The real-life mobility model: RLMM. In *Second International Conference on Future Generation Communication Technologies (FGCT 2013)*. 201–206. https://doi.org/10.1109/FGCT.2013.6767180

[13] Marco Gramaglia, Marco Fiore, Angelo Furno, and Razvan Stanica. 2021. GLOVE: Towards Privacy-Preserving Publishing of Record-Level-Truthful Mobile Phone Trajectories. *ACM/IMS Trans. Data Sci.* 2, 3, Article 21 (aug 2021), 36 pages. https://doi.org/10.1145/3451178

[14] Jie Gui, Zhenan Sun, Yonggang Wen, Dacheng Tao, and Jieping Ye. 2021. A Review on Generative Adversarial Networks: Algorithms, Theory, and Applications. *IEEE Transactions on Knowledge and Data Engineering* (2021), 1–1. https://doi.org/10.1109/TKDE.2021.3130191

[15] Ben Hughes, Shruti Bothe, Hasan Farooq, and Ali Imran. 2019. Generative Adversarial Learning for Machine Learning empowered Self Organizing 5G Networks. In *2019 International Conference on Computing, Networking and Communications (ICNC)*. 282–286. https://doi.org/10.1109/ICCNC.2019.8685527

[16] Sibren Isaacman, Richard Becker, Ramón Cáceres, Margaret Martonosi, James Rowland, Alexander Varshavsky, and Walter Willinger. 2012. Human Mobility Modeling at Metropolitan Scales. In *Proceedings of the 10th International Conference on Mobile Systems, Applications, and Services* (Low Wood Bay, Lake District, UK) *(MobiSys '12)*. Association for Computing Machinery, New York, NY, USA, 239–252. https://doi.org/10.1145/2307636.2307659

[17] Ari Jaakola, Teemu Vass, Solja Saarto, and Lotta Haglund. 2019. *Helsinki facts and figures 2019*. Technical Report. Helsinki, Finland.

[18] Shan Jiang, Gaston A Fiore, Yingxiang Yang, Joseph Ferreira Jr, Emilio Frazzoli, and Marta C González. 2013. A review of urban computing for mobile phone traces: Current methods, challenges and opportunities. In *Proceedings of the 2nd ACM SIGKDD international workshop on Urban Computing*. ACM, 2.

[19] Shan Jiang, Yingxiang Yang, Siddharth Gupta, Daniele Veneziano, Shounak Athavale, and Marta C. González. 2016. The TimeGeo modeling framework for urban mobility without travel surveys. *Proceedings of the National Academy of Sciences* 113, 37 (2016), E5370–E5378. https://doi.org/10.1073/pnas.1524261113 arXiv:https://www.pnas.org/doi/pdf/10.1073/pnas.1524261113

[20] Ari Keränen, Jörg Ott, and Teemu Kärkkäinen. 2009. The ONE Simulator for DTN Protocol Evaluation *(Simutools '09)*. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), Brussels, BEL, Article 55, 10 pages. https://doi.org/10.4108/ICST.SIMUTOOLS2009.5674

[21] Ghazaleh Khodabandelou, Vincent Gauthier, Marco Fiore, and Mounim A. El-Yacoubi. 2019. Estimation of Static and Dynamic Urban Populations with Mobile Network Metadata. *IEEE Transactions on Mobile Computing* 18, 9 (2019), 2034–2047. https://doi.org/10.1109/TMC.2018.2871156

[22] Anne Josiane Kouam, Aline Carneiro Viana, and Alain Tchana. 2021. <italic>SIMBox</italic> Bypass Frauds in Cellular Networks: Strategies, Evolution, Detection, and Future Directions. *IEEE Communications Surveys Tutorials* 23, 4 (2021), 2295–2323. https://doi.org/10.1109/COMST.2021.3100916

[23] Vaibhav Kulkarni and Benoît Garbinato. 2017. Generating Synthetic Mobility Traffic Using RNNs. In *Proceedings of the 1st Workshop on Artificial Intelligence and Deep Learning for Geographic Knowledge Discovery* (Los Angeles, California) *(GeoAI '17)*. Association for Computing Machinery, New York, NY, USA, 1–4. https://doi.org/10.1145/3149808.3149809

[24] Håvard Kvamme and Ørnulf Borgan. 2021. Continuous and discrete-time survival prediction with neural networks. *Lifetime Data Analysis* 27, 4 (oct 2021), 710–736. https://doi.org/10.1007/s10985-021-09532-6

[25] Zinan Lin, Alankar Jain, Chen Wang, Giulia Fanti, and Vyas Sekar. 2020. Using GANs for Sharing Networked Time Series Data: Challenges, Initial Promise, and Open Questions. In *Proceedings of the ACM Internet Measurement Conference* (Virtual Event, USA) *(IMC '20)*. Association for Computing Machinery, New York, NY, USA, 464–483. https://doi.org/10.1145/3419394.3423643

[26] Pablo Alvarez Lopez, Michael Behrisch, Laura Bieker-Walz, Jakob Erdmann, Yun-Pang Flötteröd, Robert Hilbrich, Leonhard Lücken, Johannes Rummel, Peter Wagner, and Evamarie Wießner. 2018. Microscopic Traffic Simulation using SUMO, In The 21st IEEE International Conference on Intelligent Transportation Systems. *IEEE Intelligent Transportation Systems Conference (ITSC)*, 2575–2582. https://doi.org/10.1109/ITSC.2018.8569938

[27] Darakhshan Mir, Sibren Isaacman, Ramon Caceres, Margaret Martonosi, and Rebecca Wright. 2013. DP-WHERE: Differentially private modeling of human mobility. *Proceedings - 2013 IEEE International Conference on Big Data, Big Data 2013*, 580–588. https://doi.org/10.1109/BigData.2013.6691626

[28] Yves-Alexandre Montjoye, Cesar Hidalgo, Michel Verleysen, and Vincent Blondel. 2013. Unique in the Crowd: The Privacy Bounds of Human Mobility. *Scientific reports* 3 (03 2013), 1376. https://doi.org/10.1038/srep01376

[29] Eduardo Mucelli Rezende Oliveira, Aline Carneiro Viana, K.P. Naveen, and Carlos Sarraute. 2017. Mobile data traffic modeling: Revealing temporal facets. *Computer Networks* 112 (2017), 176–193. https://doi.org/10.1016/j.comnet.2016.10.016

[30] Eduardo Mucelli Rezende Oliveira, Aline Carneiro Viana, Carlos Sarraute, Jorge Brea, and Ignacio Alvarez-Hamelin. 2016. On the regularity of human mobility. *Pervasive and Mobile Computing* 33 (2016), 73–90. https://doi.org/10.1016/j.pmcj.2016.04.005

[31] Aarti Munjal, Tracy Camp, and William C. Navidi. 2011. SMOOTH: A Simple Way to Model Human Mobility *(MSWiM '11)*. Association for Computing Machinery, New York, NY, USA, 351–360. https://doi.org/10.1145/2068897.2068957

[32] A. Murtić, M. Maljić, S. L. Gruičić, D. Pintar, and M. Vranić. 2018. SNA-based artificial call detail records generator. In *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*. 1226–1230. https://doi.org/10.23919/MIPRO.2018.8400222

[33] OpenCellID. 2022. The world's largest Open Database of Cell Towers. https://www.opencellid.org/

[34] OpenStreetMap contributors. 2017. Planet dump retrieved. https://www.openstreetmap.org.

[35] Metin Ozturk, Attai Ibrahim Abubakar, João Pedro Battistella Nadas, Rao Naveed Bin Rais, Sajjad Hussain, and Muhammad Ali Imran. 2021. Energy Optimization in Ultra-Dense Radio Access Networks via Traffic-Aware Cell Switching. *IEEE Transactions on Green Communications and Networking* 5, 2 (2021), 832–845. https://doi.org/10.1109/TGCN.2021.3056235

[36] Christos Papageorgiou, Konstantinos Birkos, Tasos Dagiuklas, and Stavros Kotsopoulos. 2012. Modeling Human Mobility in Obstacle-Constrained Ad Hoc Networks. *Ad Hoc Netw.* 10, 3 (may 2012), 421–434. https://doi.org/10.1016/j.adhoc.2011.07.012

[37] Luca Pappalardo and Filippo Simini. 2016. Modelling individual routines and spatio-temporal trajectories in human mobility. *CoRR* abs/1607.05952 (2016). arXiv:1607.05952 http://arxiv.org/abs/1607.05952

[38] Siyang Qin, Youchen Zuo, Yaguan Wang, Xuan Sun, and Honghui Dong. 2017. Travel trajectories analysis based on call detail record data. In *2017 29th Chinese Control And Decision Conference (CCDC)*. 7051–7056. https://doi.org/10.1109/CCDC.2017.7978454

[39] Daniel Rhoads, Ivan Serrano, Javier Borge-Holthoefer, and Albert Solé-Ribalta. 2020. Measuring and mitigating behavioural segregation using Call Detail Records. *EPJ Data Science* 9 (12 2020). https://doi.org/10.1140/epjds/s13688-020-00222-1

[40] Roselina Sallehuddin, Subariah Ibrahim, Azlan Zain, and Haashi Elmi. 2015. Detecting SIM Box Fraud by Using Support Vector Machine and Artificial Neural Network. *Jurnal Teknologi* 74 (04 2015), 137–149. https://doi.org/10.11113/jt.v74.2649

[41] Matthias Schwamborn and Nils Aschenbruck. 2013. Introducing Geographic Restrictions to the SLAW Human Mobility Model. In *2013 IEEE 21st International Symposium on Modelling, Analysis and Simulation of Computer and Telecommunication Systems*. 264–272. https://doi.org/10.1109/MASCOTS.2013.34

[42] Milita Songailaitė and Tomas Krilavičius. 2021. Synthetic call detail records generator. *CEUR Workshop proceedings [electronic resource]: IVUS 2021, proceedings of the 26th international conference on information society and university studies, Kaunas, Lithuania, April 23, 2021 / edited by Ilona Veitaitė, Audrius Lopata, Tomas Krilavičius, Marcin Woźniak. Aachen: CEUR-WS, 2021, Vol. 2915* (2021).

[43] Greta Vallero, Daniela Renga, Michela Meo, and Marco Ajmone Marsan. 2019. Greener RAN Operation Through Machine Learning. *IEEE Transactions on Network and Service Management* 16, 3 (2019), 896–908. https://doi.org/10.1109/TNSM.2019.2923881

[44] Vladimir Vukadinovic, Ólafur Ragnar Helgason, and Gunnar Karlsson. 2013. An analytical model for pedestrian content distribution in a grid of streets. *Math. Comput. Model.* 57 (2013), 2933–2944.

[45] Wikipedia. 2022. Configuration Model. https://en.wikipedia.org/wiki/Configuration_model

[46] Kai Xu, Rajkarn Singh, Hakan Bilen, Marco Fiore, Mahesh K. Marina, and Yue Wang. 2022. CartaGenie: Context-Driven Synthesis of City-Scale Mobile Network Traffic Snapshots. In *2022 IEEE International Conference on Pervasive Computing and Communications (PerCom)*. 119–129. https://doi.org/10.1109/PerCom53586.2022.9762395

[47] Kai Xu, Rajkarn Singh, Marco Fiore, Mahesh K. Marina, Hakan Bilen, Muhammad Usama, Howard Benn, and Cezary Ziemlicki. 2021. SpectraGAN: Spectrum Based Generation of City Scale Spatiotemporal Mobile Network Traffic Data. In *Proceedings of the 17th International Conference on Emerging Networking EXperiments and Technologies* (Virtual Event, Germany) *(CoNEXT '21)*. Association for Computing Machinery, New York, NY, USA, 243–258. https://doi.org/10.1145/3485983.3494844

[48] Yu Zheng, Xing Xie, and Wei-Ying Ma. 2010. GeoLife: A Collaborative Social Networking Service among User, location and trajectory. *IEEE Data(base) Engineering Bulletin* (June 2010). https://www.microsoft.com/en-us/research/publication/geolife-a-collaborative-social-networking-service-among-user-location-and-trajectory/

[49] Michael Zilske and Kai Nagel. 2014. Studying the Accuracy of Demand Generation from Mobile Phone Trajectories with Synthetic Data. *Procedia Computer Science* 32 (12 2014), 802–807. https://doi.org/10.1016/j.procs.2014.05.494