Coordinated Clustering Algorithms to Support Charging Infrastructure Design for Electric Vehicles

Marjan Momtazpour[1], Patrick Butler[1], M. Shahriar Hossain[1],
Mohammad C. Bozchalui[2], Naren Ramakrishnan[1], Ratnesh Sharma[2]
[1]Department of Computer Science, Virginia Tech, VA, 24060, USA
[2]NEC Laboratories America, Inc., CA, 95014, USA
{marjan, pabutler, msh}@cs.vt.edu,
mohammad@sv.nec-labs.com, naren@cs.vt.edu, ratnesh@sv.nec-labs.com

ABSTRACT

The confluence of several developments has created an opportune moment for energy system modernization. In the past decade, smart grids have attracted many research activities in different domains. To realize the next generation of smart grids, we must have a comprehensive understanding of interdependent networks and processes. Next-generation energy systems networks cannot be effectively designed, analyzed, and controlled in isolation from the social, economic, sensing, and control contexts in which they operate. In this paper, we develop coordinated clustering techniques to work with network models of urban environments to aid in placement of charging stations for an electrical vehicle deployment scenario. We demonstrate the multiple factors that can be simultaneously leveraged in our framework in order to achieve practical urban deployment. Our ultimate goal is to help realize sustainable energy system management in urban electrical infrastructure by modeling and analyzing networks of interactions between electric systems and urban populations.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: [Database Applications - Data mining - Spatial databases and GIS]; I.5.3 [**Pattern Recognition**]: [Clustering]; I.2.6 [**Artificial Intelligence**]: [Learning]

General Terms

Experimentation, Algorithms, Design, Measurement

Keywords

Data mining, clustering, coordinated clustering, smart grids, electric vehicles, synthetic populations.

1. INTRODUCTION

The impending decline of fossil fuels is rapidly ushering an emphasis from traditional methods of energy production, distribution, and consumption to sustainable approaches [11]. The advent of electric vehicles (EVs) is one such promising shift but to prepare for a world laden with EVs we must revisit smart grid design and operation.

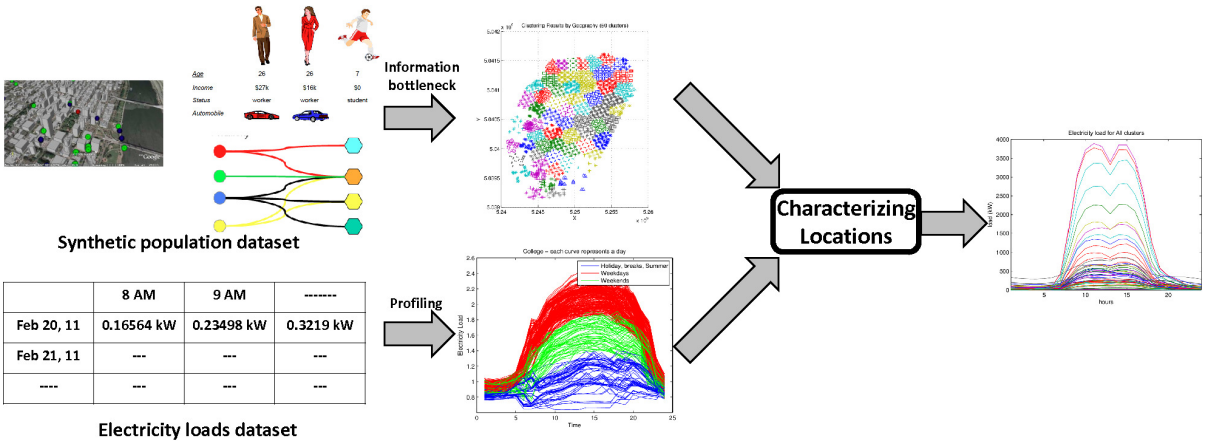
One of the key issues in ushering in EVs is the design and placement of charging infrastructure to support their operation. Issues to be taken into account include [11]: (i) prediction of EV charging needs based on their owners' activities; (ii) prediction of EV charging demands at different locations in the city, and available charge of EV batteries; (iii) design of distributed mechanisms that manage the movements of EVs to different charging stations; and (iv) optimizing the charging cycles of EVs to satisfy users' requirements, while maximizing vehicle-to-grid profits.

In this paper, we address the charging infrastructure design problem by adopting an urban computing approach. Urban computing, [8], is an emerging area which aims to foster human life in urban environments through the methods of computational science. It is focused on understanding the concepts behind events and phenomena spanning urban areas using available data sources, such as people movements and traffic flows.

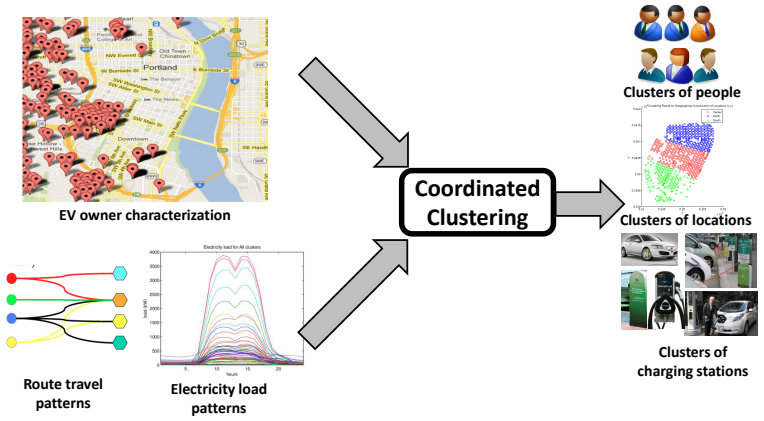
Organizing relevant data sources to solve compelling urban computing scenarios is itself an important research issue. Here, we use network datasets organized from synthetic population studies, originally designed for epidemiological scenarios, to explore the EV charging station placement problem. The dataset was organized for the SIAM Data Mining 2006 Workshop on Pandemic Preparedness [3] and models activities of an urban population in the city of Portland, Oregon. The supplied dataset [1] tracks a set of synthetic individuals in Portland and, for each of them, provides a small number of demographic attributes (age, income, work status, household structure) and daily activities representing a normative day (including places visited and times). The city itself is modeled as a set of aggregated activity locations, two per roadway link. A collection of interoperable simulations—modeling urban infrastructure, people activities, route plans, traffic, and population dynamics—mimic the time-dependent interactions of every individual in a regional area. This form of 'individual modeling' provides a bottom-up approach mirroring the contact structure of individuals and is naturally suited for formulating and studying the effect of intervention policies and considering 'what-if' scenarios.

In more detail, we characterize this dataset with a view toward understanding the behavior of EV owners and to determine which locations are most appropriate to install charging stations. We develop a coordinated clustering formulation to identify a set of locations that can be considered

ACM UrbComp '12, August 12, 2012. Beijing, China



(a) Discovering location functionalities and characterizing electricity loads.



(b) Coordinated clustering of people, locations, and charging stations.

Figure 1: Overview of our methodology.

as the best candidates for charging stations.

2. RELATED WORK

We survey related work in two categories: mining GPS datasets and smart grid analytics. GPS datasets have emerged as a popular source for modeling and mining in urban computing contexts. They have been used to extract information about roads, traffic, buildings, and people behaviors [20], [21], [9]. The range of applications is quite varied as well, from anomaly detection [9] to taxi recommender systems [21] that aim to maximize taxi-driver profits and minimize passengers' waiting times. The notion of location-aware recommender systems is a key topic enabled by the increasing availability of GPS data, e.g., recommending points of interest to tourists [22]. We survey these works in greater detail next.

In [20] Yuan et al. proposed a framework to discover regions of different functionalities based on people movements. They adapt algorithms from the topic modeling literature, by mapping a region as a document and a function as a topic so that human movements become 'words' in this model. The focus of [21] and [19] is different: here, GPS data is used to mine the fastest driving routes for taxi drivers. In [21], Yuan et al. mined smart driving direction from GPS trajectory of taxis, and in [19] they consider driver behavior using other metrics such as driving strategies and weather conditions.

Clusters of moving objects in a noisy stadium environment are detected using the DBSCAN algorithm [5] in [12]. This task supports monitoring a stadium for groups of individuals that exhibit concerted behavior. In [14], the authors estimate distributions of travel-time from GPS data for use in routing and route-recommendation.

Our work here is different from the above works in that we use a synthetic population dataset and routes are based on people's travel habits that are mapped using geographical coordinates and road infrastructures. We are also not *per se* interested in mining the routes but to use the route information to better support charging infrastructure placement.

Smart grid analytics has emerged as a promising approach to usher in the promise of smart grid benefits. Researchers have begun to explore the problems concomitant with EV penetration in urban areas, especially unacceptable increases in electricity consumption [11]. A promising way to approach this problem is to understand the interactions between grid infrastructure and urban populations. While smart grids and EVs have been studied previously from technical and AI point of views, there is a limited number of research on smart grids from an urban computing perspective.

In this space, agent-based systems have been proposed to simulate city behavior in terms of agents with a view toward designing decentralized systems and maximizing grid profits as well as individuals' profit [11]. In [2] information from smart meters is used for forecasting energy consumption patterns in a university campus micro-grid, whose re-

sults can be used for future energy planning. In our work, we directly study the problem of charging station placement using coordinated clustering algorithms.

3. METHODOLOGY

Our overall methodology is given in Figure 1. We describe each of the steps in our approach next. At a basic level, we integrate two basic types of data to formulate our data mining scenario. The first data, as described earlier, is a synthetic population of people and activities representing the city of Portland and the second data set is electricity consumption profile of each location. Notice that the proposed methodology is a generic approach and can be applied to real-world data and the fact that we use synthetic data here is only due to our lack of access to real-world data to test our proposed methodology.

The synthetic dataset contains 243,423 locations of which 1,779 are selected as belonging to the downtown area and of further interest for our purposes. Each location is represented by geographical [x,y] coordinate adopting the universal transverse mercator coordinate system (UTM) [1]. There are a total of 1,615,860 people in the entire city. Information about them is organized into households, and for each household we have the details of number of people in the household, and the ages, genders, and incomes of each household member. Each person has a unique ID.

We have some information about each person including age, gender, income, and his/her house ID. The typical movement patterns of people in a typical day (27 hour period) are also available. A total of 8,922,359 movements are provided. In addition to starting and ending locations for people's movements, this dataset also provides the *purpose* of the movement, categorized into nine types: {Home, Work, Shop, Visit, Social/Recreational, Serve Passenger, School, College, and Other}. A given person moves from one location to another location at a specific time for a specific purpose (from the nine mentioned above) and stays in that location for a specified period of time. These movement types can thus be utilized for further detailed studies. We also have the ability to map the locations using Google Maps and calculate distances of traveling between locations.

To this dataset, we augment information about electricity consumption of each location and simulate the effects of EVs on its electricity demand profile. Since actual electricity consumption data for each location is not available until all the consumers have smart meters installed and in operation for some time, we approximate electricity load profile using the existing data (organized by NEC Labs, America).

It is clear that the electricity load of each location greatly depends on the functionality of that location and hence our first approach is to utilize an information bottleneck type approach [17] to characterize locations. Our aim is to cluster locations based on geographical proximity but such that the resulting clusters are highly informative of location function. This is thus our first application of a coordinated clustering formulation, and falls in the scope of clustering with side information. Next, we integrate the electricity load information to characterize usage patterns across clusters with a view toward helping identifying locations to place charging infrastructure.

Our next step is to more accurately characterize usage patterns of likely EV owners. A specific set of clusters from the previous pipeline is used and characterized using high-income attributes as the likely owners of EVs. We then bring in additional factors of locations that influence EV charger placement, e.g., residentiality ratio, load on the location, charging needs, and typical duration of stay in the location. Some of these factors (such as distance traveled) are in turn determined by mapping the home-to-work and work-to-home trajectories of EV owners and their stop locations. We use a coordinated clustering formulation to simultaneously cluster three datasets in a relational setting. Our coordinated clustering framework builds upon our previous work [7] which generalizes relational clustering between two non-homogeneous datasets. This problem is a bit non-trivial since one of the relations is a many-to-many relation and another is a one-to-one relation. The final set of coordinated clusters are then used as interpretation and as a guide to charger placement.

After locating the homes of EV owners, we can determine their trajectories and their stop locations. Then, based on this data, we can estimate their travel distances. This helps us to estimate charging requirements of EVs, during a day. With the help of the distribution of electricity load in the city and charging needs of EVs, we determine proper locations for installing charging stations in city with respect to specific parameters.

4. ALGORITHMS

As described above, our methodology comprises the following four major steps to determine candidate locations for charging stations: (i) discovering locations' functionalities using an information bottleneck method; (ii) electricity load estimation and integrating with results of (i); (iii) studying the behavior of EV owners and calculating specific parameters relevant to their usage patterns; and (iv) candidate selection for charging stations using coordinated clustering techniques. Each of these steps are detailed next.

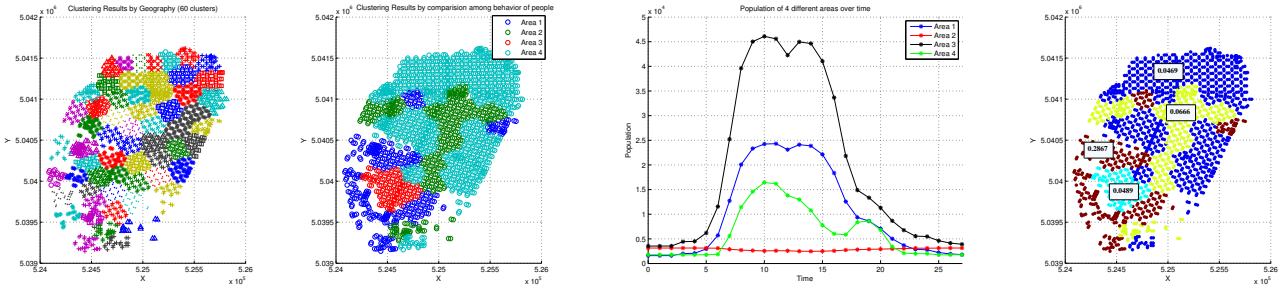
4.1 Discovering Location Functionalities

We use information bottleneck methods to characterize locations with a view toward defining the specific purpose of the location. The idea of information bottleneck methods is to cluster data points in a space (here, geography) such that the resulting clusters are highly informative of another random variable (here, function). We focus on 1779 locations in the downtown Portland area whose geographies are defined by (x,y) coordinates and whose functions are given by a 9-length profile vector $P = [p_1, p_2, ..., p_9]$, where p_i is the number of travels incident on that location for the i^{th} purpose (recall the different purposes introduced in the previous section).

Figure 2 (a) describes the results of a clustering based on euclidean metrics between locations whose results are aggregated in Figure 2 (b) into a revised clustering that also preserves information about activities of people at these locations. The population distribution of these clusters over time is shown in Figure 2 (c) which reveals characteristic changes of crowds around peak hours and lunch times. One final analysis that will be useful is to evaluate each of the discovered clusters with respect to what we term as the *residentiality ratio*. The residentiality ratio for a location is the percentage of people who use that location as a home w.r.t. all people who visit that location (in downtown Portland, many locations have combined home-work profiles, and hence the calculation of residentiality ratio becomes relevant). Figure 2 (d) reveals one cluster with relatively high residentiality ratio among three others.

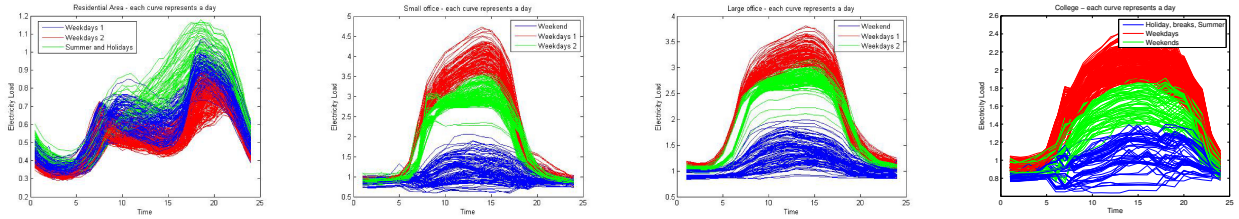
4.2 Electricity Load Estimation

In order to uncover patterns in electricity load distributions, we now characterize each of the discovered clusters us-



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

Figure 2: (a) Clustering downtown locations based on geographic coordinates. (b) Clustering over the previous clustering with people's activities as side-information. (c) Dynamic population of the four discovered clusters over a typical day. (d) Computed residentiality ratio revealing one primary residential cluster.



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

Figure 3: (a) Electricity usage in residential areas. (b) Electricity usage in small office areas. (c) Electricity usage in large office areas. (d) Electricity usage in college areas.

ing typical profiles gathered from public data sources such as the California End User Survey (CEUS) and other sources of usage information. Figure 3 presents daily electricity consumption profile across large offices, small offices, residential buildings, and colleges for one year. By clustering this data across the year, we can discern important patterns associated with different types of consumption during the year. For instance, in the college setting, we can discern three types of consumption patterns: holiday breaks (including summer), weekdays, and weekends.

Our next step is to compute the electricity load leveraging the above patterns but w.r.t. our network model of the urban environment. Recall that our network model is based on population dynamics but typical electricity load sources are based on square footage calculations. We map these factors using well-accepted measures, i.e., by considering the average square footage occupied by one person in a residential area as 600sft [4], small office as 200sft [18], large office as 200sft [18], college as 50sft [15], retail area as 50sft [15], and other classes as 200. Further, the minimum population for an office to be considered as a large office is set to 300.

Based on some exploratory data analysis, we selected a weekday in the past year (specifically, 18th March, 2011) and used the electricity load data of this day to map to the network model. Consider that in a specific hour, N people go to location l in which n_i of them come for the purpose of p_i while $\sum_{i=1}^{9} n_i = N$. Then the electricity load for that location is computed as

$$E_l = \sum_{i=1}^{9} \frac{n_i A_{p_i} E_{p_i}}{1000}, \quad (1)$$

where A is the average square footage per person and E_p is electricity consumption of building type p. Observe that a single location can serve multiple purposes and the above equation marginalizes across all uses. For example, if there are 360 people in one location, and 10 of them are in the building for the purpose of home and 350 are for the purpose

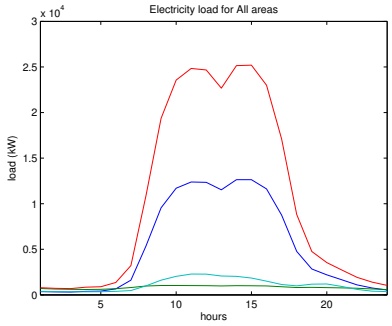


Figure 4: Electricity loads for four characterized location clusters.

of office, the total electricity consumption of building would be calculated as $(10 \times 600 \times E_{p_{\text{home}}}/1000) + (350 \times 200 \times E_{p_{\text{office}}}/1000)$ where 600 and 200 are average square footage per person for the different categories, as mentioned earlier. The above methodology enables us to characterize electricity loads in terms of the four location clusters characterized in the previous step (see Figure 4).

4.3 Characterizing EV users

Currently only a small percentage of people use EVs, and this figure is correlated with high income. Based on [10] and [13], only 6 percent of people in the US have income more than 170,000 USD. In our synthetic dataset, 329,218 people make an income greater than 60,000 USD. To explore a hypothetical scenario, we posed the question:

> What if 6.31% of 329,218 people from Portland bought EVs? What charging infrastructure is necessary to support this scenario?

Based on our modeling of these people's movements and patterns, we aim to identify the best locations for charging stations.

Figure 5 (a) gives the distribution of EV users in our potential scenario. We can notice several clusters around high-income neighborhoods. With the aid of Google Maps, we can estimate the amount of time an EV owner drives and how far he/she travels on a regular week day. Figure 5 (b) gives the distribution of distances traveled by these users.

Assuming EV owners charge their cars at their respective homes, our goal is now to identify candidate charging locations during other times. Let us assume that the EV of a person P consumes E_P^C KWh energy per 100 Km. Also, assume that the battery of this vehicle can save E_P^S KWh. Then the estimated total distance that P can travel with his vehicle before he needs to charge its battery is

$$\Delta_P = \frac{100 E_P^S}{E_P^C}, \tag{2}$$

As an example, for the Chevrolet Volt [6], with $E_P^S = 16$ KWh and $E_P^C = 22.4$ KWh per 100 Km, the EV can travel 71.43 Km before it needs to be recharged.

If the total traveling distance of P in a day is D_P then the number of times that P needs to charge his vehicle is N_P and is determined as follows:

$$N_P = \left\lfloor \frac{D_P}{\Delta_P} \right\rfloor, \tag{3}$$

As an example, if we assume that an EV's battery can save 16 KWh energy [6], an electric car can go for 71.43 Km before it needs to be charged [16].

Due to the long duration of charging process, we have a constraint to install charging stations only in destinations that people visit. Assume that V_L is the set of EV owners who visited location L during the day. Then $|V_L|$ is the total number of EV owners who have visited location L. However, there is a greater chance for a location to be a charging station if people with higher charge needs visit that location. Hence, the charge needs of location L is determined based on equation 4.

$$W_L = \sum_{P \in V_L} N_P, \tag{4}$$

Figure 5 (c) depicts the histogram of how many times an EV needs to be charged. Also, Figure 5 (d) depicts the charge needs of downtown locations.

On the other hand, each person visit a location for a specific period of time which here we call it duration of stay. In order to put a charging station in one location, we force people to stay for a specific period of time because charging an electric car will take couple of hours. Hence, in locations where people stay longer such as working locations have potential to be charging stations compared to those locations that people stay in them for example half an hour. We use average duration of stay of people in each location as a feature for that location.

It should be noted that the right choice of EV charging stations depends on regular electricity load of each area, the amount of time that each person spends on this location, and number of times that EV owners need to charge their vehicles. Hence, based on EV owners traveling route during peak and off-peak hours, we can come up with a set of candidate regions for charging stations.

4.4 Charging Station Placement using Coordinated Clustering

Since charging EVs is not an instantaneous process, it is helpful to place charging stations at those locations where people visit for an extended period of time. The average duration of stay of people in each location is an important feature in this regard. The right choice of EV charging stations thus depends on the regular electricity load of the area, the amount of time that people spend in the location, and the number of times that EV owners need to charge their vehicles. Hence, based on EV owners' traveling routes during peak and off-peak hours, we can arrive at a set of candidate regions for charging stations.

Let $\mathcal{X}$ be the income dataset and $\mathcal{Y}$ be the locations datasets. $\mathcal{X} = \{\mathbf{x}_s\}, s = 1, \ldots, n_x$ is the set of vectors in dataset $\mathcal{X}$, where each vector is of dimension l_x, i.e., $\mathbf{x}_s \in \mathbb{R}^{l_x}$. Currently, our income dataset contains only one dimension. Similarly, locations dataset $\mathcal{Y} = \{\mathbf{y}_t\}, t = 1, \ldots, n_y, \mathbf{y}_t \in \mathbb{R}^{l_y}$. Locations are denoted by two dimensions (latitude and longitude) in our current database. The many-to-many relationships between $\mathcal{X}$ and $\mathcal{Y}$ are represented by a $n_x \times n_y$ binary matrix B, where $B(s,t) = 1$ if $\mathbf{x}_s$ is related to $\mathbf{y}_t$, else $B(s,t) = 0$. Let $C_{(x)}$ and $C_{(y)}$ be the cluster indices, i.e., indicator random variables, corresponding to the income dataset $\mathcal{X}$ and location dataset $\mathcal{Y}$ and let k_x and k_y be the corresponding number of clusters. Thus, $C_{(x)}$ takes values in $\{1, \ldots, k_x\}$ and $C_{(y)}$ takes values in $\{1, \ldots, k_y\}$.

Let $\mathbf{m}_{i,\mathcal{X}}$ be the prototype vector for cluster i in income dataset $\mathcal{X}$ (similarly $\mathbf{m}_{j,\mathcal{Y}}$). These are the variables we wish to estimate/optimize for. Let $v_i^{(\mathbf{x}_s)}$ (likewise $v_j^{(\mathbf{y}_t)}$) be the cluster membership indicator variables, i.e., the probability that income data sample $\mathbf{x}_s$ is assigned to cluster i in the income dataset $\mathcal{X}$ (resp). Thus, $\sum_{i=1}^{r^x} v_i^{(\mathbf{x}_s)} = \sum_{j=1}^{r^y} v_j^{(\mathbf{y}_t)} = 1$. The traditional k-means *hard* assignment is given by:

$$v_i^{(\mathbf{x}_s)} = \begin{cases} 1 & \text{if } ||\mathbf{x}_s - \mathbf{m}_{i,\mathcal{X}}|| \leq ||\mathbf{x}_s - \mathbf{m}_{i',\mathcal{X}}||,\ i' = 1 \ldots k_x, \\ 0 & \text{otherwise.} \end{cases}$$

(Likewise for $v_j^{(\mathbf{y}_t)}$.) Ideally, we would like a continuous function that tracks these hard assignments to a high degree of accuracy. Such a continuous function for the the cluster membership can be defined as follows.

$$v_i^{(\mathbf{x}_s)} = \frac{\exp(-\frac{\rho}{D}||\mathbf{x}_s - \mathbf{m}_{i,\mathcal{X}}||^2)}{\sum_{i'=1}^{k_x} \exp(-\frac{\rho}{D}||\mathbf{x}_s - \mathbf{m}_{i',\mathcal{X}}||^2)} \tag{5}$$

where ρ is a user-settable parameter and D is the pointset diameter which depends on the data.

An analogous equation holds for $v_j^{(\mathbf{y}_t)}$.

We prepare a $k_x \times k_y$ contingency table to capture the relationships between entries in clusters across income dataset $\mathcal{X}$ and locations dataset $\mathcal{Y}$. To construct this contingency table, we simply iterate over every combination of data entities from $\mathcal{X}$ and $\mathcal{Y}$, determine whether they have a relationship, and suitably increment the appropriate entry in the contingency table:

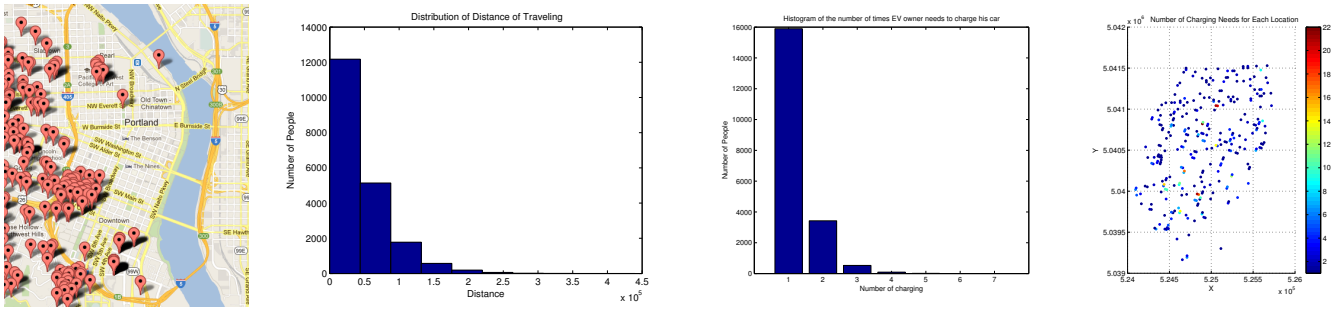
$$w_{ij} = \sum_{s=1}^{n_x} \sum_{t=1}^{n_y} B(s,t) v_i^{(\mathbf{x}_s)} v_j^{(\mathbf{y}_t)}. \tag{6}$$

We also define

$$w_{i.} = \sum_{j=1}^{k_y} w_{ij}, \qquad w_{.j} = \sum_{i=1}^{k_x} w_{ij},$$

where $w_{i.}$ and $w_{.j}$ are the row-wise (income cluster-wise) and column-wise (locations cluster-wise) counts of the cells of the contingency table respectively.

We also define the row-wise random variables $\alpha_i, i = 1, \ldots, k_x$ and column-wise random variables $\beta_j, j = 1, \ldots, k_y$ with



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

Figure 5: (a) EV household locations. (b) Distribution of distances people travel in their EVs. (c) Charging needs for EVs. (d) Number of charging needs (more than zero) per location.

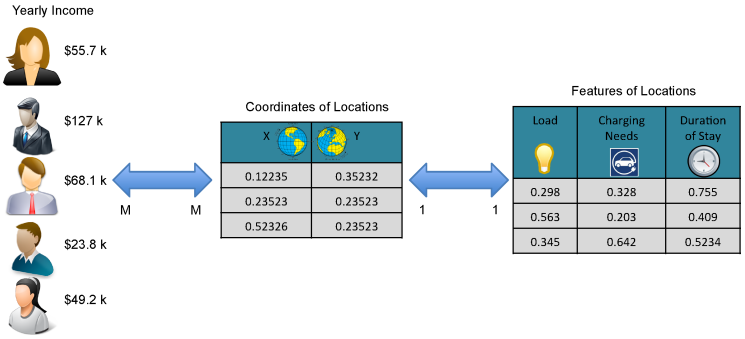


Figure 6: Coordinated clustering schema.

probability distributions as follows

$$p(\alpha_i = j) = p(C_{(y)} = j | C_{(x)} = i) = \frac{w_{ij}}{w_{i.}}. \quad (7)$$

$$p(\beta_j = i) = p(C_{(x)} = i | C_{(y)} = j) = \frac{w_{ij}}{w_{.j}}. \quad (8)$$

The row-wise distributions represent the conditional distributions of the clusters in dataset in $\mathcal{X}$ given the clusters in $\mathcal{Y}$; the column-wise distributions are also interpreted analogously.

After we construct the contingency table, we must evaluate it to see if it reflects a coordinated clustering. In coordinated clustering, we expect that the contingency table will be nonuniform. We can expect that the contingency table will be an identity matrix when $k_x = k_y$. To keep the formulation and the implementation generic for different number of clusters in two dataset, we need to optimize the variables (cluster prototypes) in such a way that the contingency table is far from its uniform case. For this purpose, we compare the income cluster (row-wise) and locations cluster (column-wise) distributions from the contingency table entries to the uniform distribution.

We use KL-divergences to define our unified objective function:

$$\mathcal{F} = \frac{1}{k_x}\sum_{i=1}^{k_x} D_{KL}\left(\alpha_i || U\left(\frac{1}{k_y}\right)\right) + \frac{1}{k_y}\sum_{j=1}^{k_y} D_{KL}\left(\beta_j || U\left(\frac{1}{k_x}\right)\right), \quad (9)$$

where D_{KL} is the KL-divergence between two distributions and U indicates the uniform distribution over a row or a column.

Note that the row-wise distributions take values over the columns $1, \ldots, k_y$ and the column-wise distributions take values over the rows $1, \ldots, k_x$. Hence the reference distribution for row-wise variables is over the columns, and vice versa. Also, observe that the row-wise and column-wise KL-divergences are averaged to form $\mathcal{F}$. This is to mitigate the effect of lopsided contingency tables ($k_x \gg k_y$ or $k_y \gg k_x$) wherein it is possible to optimize $\mathcal{F}$ by focusing on the "longer" dimension without really ensuring that the other dimension's projections are close to uniform.

Maximizing $\mathcal{F}$ leads to rows (income clusters) and columns (locations clusters) in the contingency table that are far from the uniform distribution as required by the coordinated clusters. It is equivalent to minimizing $-\mathcal{F}$.

The coordinated clustering formulation presented thus far can have some degenerate solutions where large number of data points in both datasets are assigned to the same cluster leading to a huge overlap of relationships. To mitigate this, we add two more terms with the objective function.

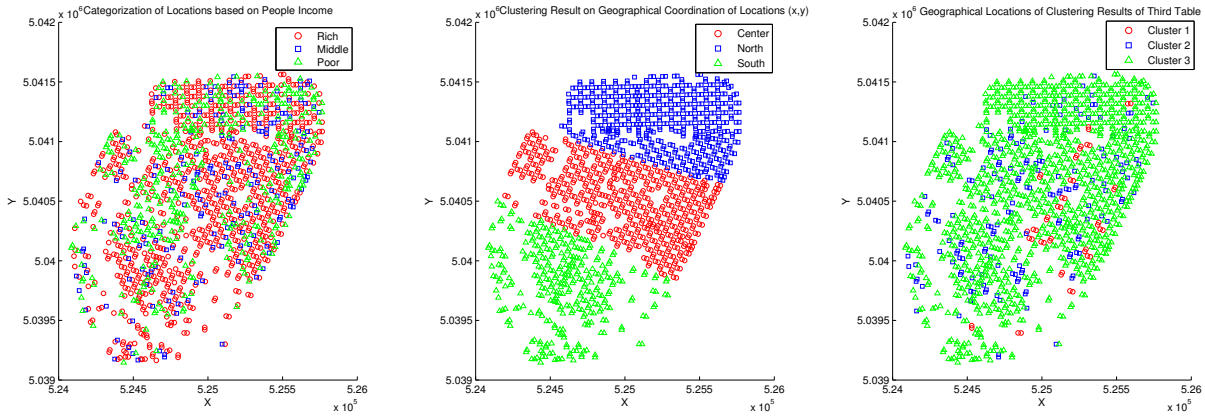
$$\mathcal{F}_{\mathcal{R}} = -\mathcal{F} + D_{KL}\left(p(\alpha) || U\left(\frac{1}{k_x}\right)\right) + D_{KL}\left(p(\beta) || U\left(\frac{1}{k_y}\right)\right). \quad (10)$$

It should be noted that function $\mathcal{F}_{\mathcal{R}}$ is expected to be minimized. This is the reason why $-\mathcal{F}$ is used in the formula for $\mathcal{F}_{\mathcal{R}}$.

Finally, we describe how to integrate three datasets: income, location, and station properties. Let $\mathcal{X}$, $\mathcal{Y}$, and $\mathcal{Z}$ be these three datasets, respectively. There are two sets of relationships, existing between $\mathcal{X}$, $\mathcal{Y}$, and $\mathcal{Y}$, $\mathcal{Z}$. The objective function for these three datasets and two sets of relationships is defined as follows.

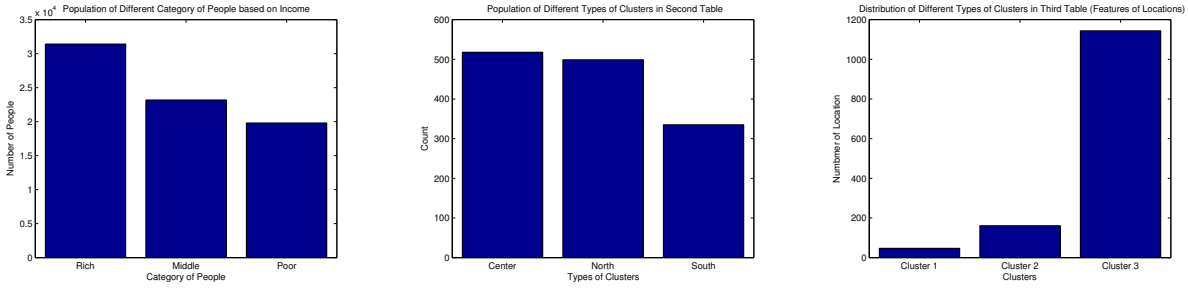
$$\mathcal{F}_{\mathcal{XYZ}} = \mathcal{F}_{\mathcal{R}}(\mathcal{X}, \mathcal{Y}) + \mathcal{F}_{\mathcal{R}}(\mathcal{Y}, \mathcal{Z}). \quad (11)$$

Here $\mathcal{F}_{\mathcal{R}}(\mathcal{X}, \mathcal{Y})$ refers to the objective function described in Eq. 10 with the income dataset $\mathcal{X}$, and locations dataset $\mathcal{Y}$. $\mathcal{F}_{\mathcal{R}}(\mathcal{Y}, \mathcal{Z})$ refers to the same objective function but input datasets are locations $\mathcal{Y}$, and station property $\mathcal{Z}$. In all



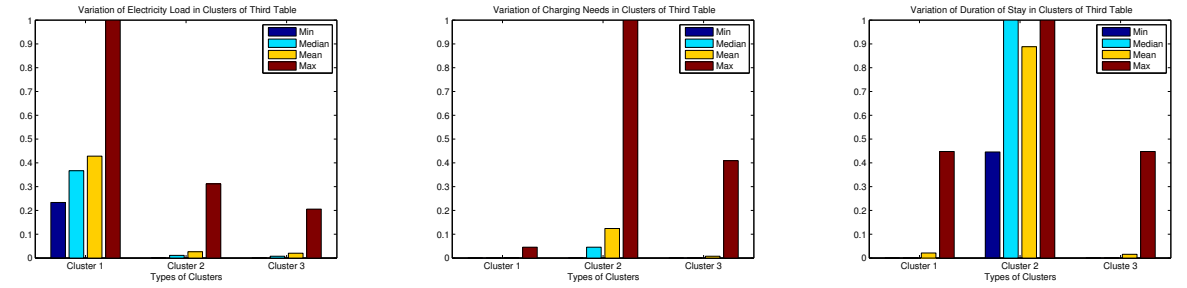
(a) (b) (c)

Figure 7: Results of coordinated clustering (3 clusters) when viewed through the attributes of each domain. (a) Clusters based on income. (b) Clusters based on geographical location. (c) Clusters based on EV charging station attributes.



(a) (b) (c)

Figure 8: Profiles of clusters obtained from coordinated clustering w.r.t. each of the three domains. (a) Income attributes. (b) Location attributes. (c) EV charging station attributions.



(a) (b) (c)

Figure 9: Detailed inspection of clusters for their suitability for locating EV charging stations. (a) Distribution of electricity loads. (b) Distribution of charging needs. (c) Distribution of duration of stay. An ideal cluster should have (low, high, high) values respectively, suggesting that cluster 2 is best suited.

our experiments, we minimize $\mathcal{F}_{\mathcal{X}\mathcal{Y}\mathcal{Z}}$ to apply coordinated clustering between income, locations, and station property datasets.

5. RESULTS

Figure 6 describes the coordinated clustering scenario involving: yearly income, a location's geographical coordinates, and the location's features.

We begin with some preliminary observations about our data. Figure 10 depicts the distribution of people based on their income, indicating that a significant number of people have high income, leading to a large number of EV users. We experimented with coordinated clustering settings involving many settings. Figure 7 depicts three clusters of locations based on each of the attribute sets in our schema. Note that because the clusters are mapped onto (x,y) geographical locations, locality is apparent only in Figure 7 (b).

Profiles of these clusters are described in detail in Figure 8. Of particular interest to us is the view from the perspective of EV attributes, i.e., Figure 8 (c). Details of these clusters are explored in greater detail in Table. 1. Ideal locations for charging stations for EVs must have a relatively low current electricity load (to accommodate the installation of charging infrastructure), high charging needs (population profiles), and high staying duration. As can be seen from Table. 1 cluster 2 fits these requirements. Greater insights into the three clusters from the viewpoint of these

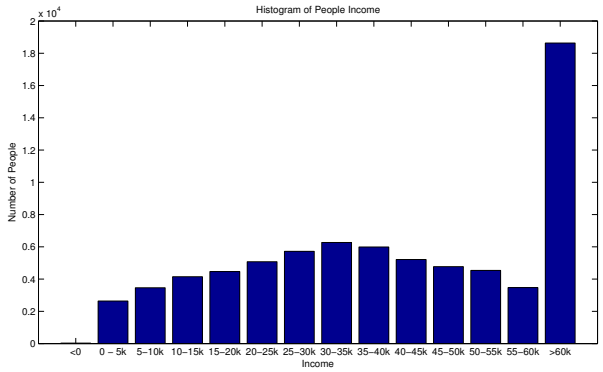


Figure 10: Distribution of income.

Table 1: Characteristics of Clusters in Third Table (Location's Features)

Cluster	Elec. Load	Charging Need	Stay Duration
1	High	Low	Low
2	Low	High	High
3	Low	Low	Low

three attributes is shown in Figure 9 supporting the choice of locations in cluster 2 as the right candidates for locating charging stations.

6. DISCUSSION

Electrical vehicles are only going to become more popular in the near future. We have demonstrated a systematic data mining methodology that can be used to identify locations for placing charging infrastructure as EV needs grow. The results presented here can be generalized to a temporal scenario where we accommodate a growing EV population and to design charging infrastructure to accommodate additional scenarios of smart grid usage and design.

The methodology presented in this paper only incorporates demand data from the electricity infrastructure and future work would incorporate information from the electricity supply side too. Information such loading level of electricity feeders and remaining excess capacity of feeders for EV charging stations can be integrated in presented methodology to improve the placement of EV charging stations. Moreover, this methodology can be used to identify locations of interest for deployment of stationary energy storages to more efficiently utilize existing electricity infrastructure rather than building new expensive transmission capacity to meet the demand of EV charging stations.

7. REFERENCES

[1] Synthetic Data Products for Societal Infrastructures and Proto-Populations: Data Set 1.0. Technical Report NDSSL-TR-06-006, Network Dynamics and Simulation Science Laboratory, Virginia Tech, Blacksburg, VA, 2006.

[2] S. Aman, Y. Simmhan, and V. K. Prasanna. Improving Energy Use Forecast for Campus Micro-grids using Indirect Indicators. In *IEEE Workshop on Domain Driven Data Mining*, 2011.

[3] C. Bailey-Kellogg, N. Ramakrishnan, and M. Marathe. Spatial data mining to support pandemic preparedness. *SIGKDD Explor. Newsl.*, 8(1):80–82, Jun 2006.

[4] K. S. Blake, R. L. Kellerson, and A. Simic. Measuring Overcrowding in Housing, September 2007.

[5] M. Ester, H. Kriegel, J. Sander, and X. Xu. A Density-based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *KDD '96*, pages 226–231, 1996.

[6] GM-Volt. Chevrolet Volt Specifications. Last accessed May 16 2012 `http://Gm-volt.com/full-specification`.

[7] M. S. Hossain, S. Tadepalli, L. T. Watson, I. Davidson, R. F. Helm, and N. Ramakrishnan. Unifying Dependent Clustering and Disparate Clustering for Non-homogeneous Data. In *KDD '10*, pages 593–602, 2010.

[8] T. Kindberg, M. Chalmers, and E. Paulos. Urban Computing. *IEEE Pervasive Computing*, pages 18–20, July-September 2007.

[9] W. Liu, Y. Zheng, S. Chawla, J. Yuan, and X. Xie. Discovering Spatio-Temporal Causal Interactions in Traffic Data Streams. In *KDD '11*, August 2011.

[10] N. Munro. Obama Hikes Subsidy to Wealthy Electric Car Buyers. Last accessed May 16 2012 `http://dailycaller.com/2012/02/13/obama-hikes-subsidy-to-wealthy-electric-car-buyers/`.

[11] S. D. Ramchurn, P. Vytelingum, A. Rogers, and N. R. Jennings. Putting the 'Smarts' into the Smart Grid: A Grand Challenge for Artificial Intelligence. *Communications of the ACM*, 55(4):86–97, April 2012.

[12] J. Rosswog and K. Ghose. Detecting and Tracking Coordinated Groups in Dense, Systematically Moving, Crowds. In *SDM '12*, pages 1–11, 2012.

[13] Simply Hired, Inc. Portland Jobs. Last accessed May 16 2012 `http://www.simplyhired.com/a/local-jobs/city/l-Portland,+OR`.

[14] R. Takahashi, T. Osogami, and T. Morimura. Large-Scale Nonparametric Estimation of Vehicle Travel Time Distributions. In *SDM '12*, pages 12–23, April 2012.

[15] The Engineering ToolBox. Common Area per Person in Buildings. Last accessed: May 16 2012 `http://www.engineeringtoolbox.com/number-persons-buildings-d_118.html`.

[16] The official U.S. Government Source for Fuel Economy Information. Fuel Economy. Last accessed May 16 2012 `http://www.fueleconomy.gov/feg/`,.

[17] N. Tishby, F. C. Pereira, and W. Bialek. The Information Bottleneck Method. In *37th Annual Allerton Conference on Communication, Control and Computing*, pages 368–377, 1999.

[18] U.S. General Services Administration. Office Space Use Review, Current Practices and Emerging Trends, September 1997.

[19] J. Yuan, Y. Zheng, and et al. Driving with Knowledge from the Physical World. In *KDD '11*, 2011.

[20] J. Yuan, Y. Zheng, and X. Xie. Discovering Region of Different Functions in a City Using Human Mobility and POI. In *KDD '12*, 2012.

[21] J. Yuan, Y. Zheng, C. Zhang, W. Xie, X. Xie, and Y. Huang. T-Drive: Driving Directions Based on Taxi Trajectories. In *ACM SIGSPATIAL GIS 2010*, 2010.

[22] Y. Zheng, L. Zhang, X. Xie, and W. Ma. Mining Correlation between Locations Using Human Location History. In *ACM SIGSPATIAL GIS 2009*, 2009.