A

Charging and Storage Infrastructure Design for Electric Vehicles

MARJAN MOMTAZPOUR and PATRICK BUTLER, Virginia Tech
M. SHAHRIAR HOSSAIN, University of Texas at El Paso
MOHAMMAD C. BOZCHALUI and RATNESH SHARMA, NEC Laboratories America, Inc.
NAREN RAMAKRISHNAN, Virginia Tech

Ushered by recent developments in various areas of science and technology, modern energy systems are going to be an inevitable part of our societies. Smart grids are one of these modern systems that have attracted many research activities in recent years. Before utilizing the next generation of smart grids, we should have a comprehensive understanding of the interdependent energy 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 present a novel framework to support charging and storage infrastructure design for electric vehicles. 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. Furthermore, we evaluate the network before and after the deployment of charging stations, to recommend the installation of appropriate storage units to overcome the extra load imposed on the network by the charging stations. 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

Additional Key Words and Phrases: Clustering, coordinated clustering, data mining, electric vehicles, smart grids, storage, charging stations, synthetic populations.

ACM Reference Format:
Momtazpour, M., Butler, P., Hossain, M. S., Bozchalui, M. C., Sharma R., and Ramakrishnan, N. 2012. Charging and Storage Infrastructure Design for EVs. ACM Trans. Intell. Syst. Technol. V, N, Article A (January YYYY), 27 pages.
DOI = 10.1145/0000000.0000000 http://doi.acm.org/10.1145/0000000.0000000

1. INTRODUCTION

Due to the fast decline of fossil fuels, sustainable approaches of energy production, distribution, and consumption are now going to take the place of traditional methods [Ramchurn et al. 2012]. The advent of electric vehicles (EVs) is a promising shift. However, to be prepared for a world laden with EVs we must revisit smart grid design

This work is supported by the NEC Laboratories America, Inc. (NEC Labs).
Author's addresses: M. Momtazpour and P. Butler and N. Ramakrishnan, Department of Computer Science, Virginia Tech; M.S. Hossain, Department of Computer Science, University of Texas at El Paso; R. Sharma and M. C. Bozchalui, NEC Laboratories America, Inc., CA.

DOI 10.1145/0000000.0000000 http://doi.acm.org/10.1145/0000000.0000000

and operation. One of the key issues in ushering in EVs is the design and placement of charging infrastructure to support their operation. Issues that must be addressed include [Ramchurn et al. 2012]:

(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 propose a new framework to address the problem of charging and storage infrastructure design for EVs by adopting an urban computing approach. Furthermore, due to the additional load imposed to the network by EVs, appropriate storage units must be deployed beside the charging stations. There are several works that consider the problem of load management for EV charging and its impact on the grid [Paul and Aisu 2012] and [Guiseppe and Antonio 2012]. However, there is no previous work that addresses the coordinated impact of placement over an urban infrastructure and its solution thereof.

Here, we assume that each charging station uses storage to offset the impact of charging on the grid. Other alternative solutions can also be used, such as upgrading transmission lines or using a Vehicle-to-Grid (V2G) strategy. While we can upgrade the transmission lines, the distribution infrastructure still remains a bottleneck. In fact, upgrading transmission lines is not a complete solution, albeit expensive and time consuming, discounting the regulatory hurdles. The other solution is based on using V2G. However, there are no EVs today which provide V2G capability to owners and business models utilizing such capability are still uncertain from a utility perspective. Hence, this will not affect the proposed methodology.

Urban computing, [Kindberg et al. 2007], 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 [Bailey-Kellogg et al. 2006] and models activities of an urban population in the city of Portland, Oregon. The supplied dataset [Bisset et al. 2006] 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 our previous work [Momtazpour et al. 2012], we characterized 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 developed a coordinated clustering formulation to identify a set of locations that can be considered as the best

candidates for charging stations. However, thorough study of this problem needs an approach to determine economic costs imposed on EV owners, and to evaluate the extra load which is imposed on the network by charging stations. In the current paper, we extend our previous framework to consider charging costs and storage placement problems in addition to the problem of charging station placement. We develop an algorithm to assign EVs to the nearest charging stations by minimizing charging cost and travel distance. After assigning charging stations to EVs, additional load of each charging station is calculated and used to determine appropriate storage deployment for each location.

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 [Yuan et al. 2012], [Yuan et al. 2010], [Liu et al. 2011]. The range of applications is quite varied as well, from anomaly detection [Liu et al. 2011] to taxi recommender systems [Yuan et al. 2010] 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 [Zheng et al. 2009]. We survey these works in greater detail next.

In [Yuan et al. 2012] 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 [Yuan et al. 2010] and [Yuan et al. 2011] is different: here, GPS data is used to mine the fastest driving routes for taxi drivers. In [Yuan et al. 2010], Yuan et al. mined smart driving direction from GPS trajectory of taxis, and in [Yuan et al. 2011] 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 [Ester et al. 1996] in [Rosswog and Ghose 2012]. This task supports monitoring a stadium for groups of individuals that exhibit concerted behavior. In [Takahashi et al. 2012], 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 [Ramchurn et al. 2012]. 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 [Ramchurn et al. 2012]. In [Aman et al. 2011] information from smart meters is used for forecasting energy consumption patterns in a university campus micro-grid, whose results can be used for future energy planning.

Significant research has been done to improve cost and reliability of energy storage systems [Hoffman et al. 2010]. Energy storage is used to perform an operation when there is not enough electricity. In [Makarov et al. 2012] a solution is proposed to balance energy production against its consumption. In addition, authors in [Bayram et al. 2011] try to design a general architecture in smart grid to have a significant gains in net cost/profit considering Electric Vehicles.

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.

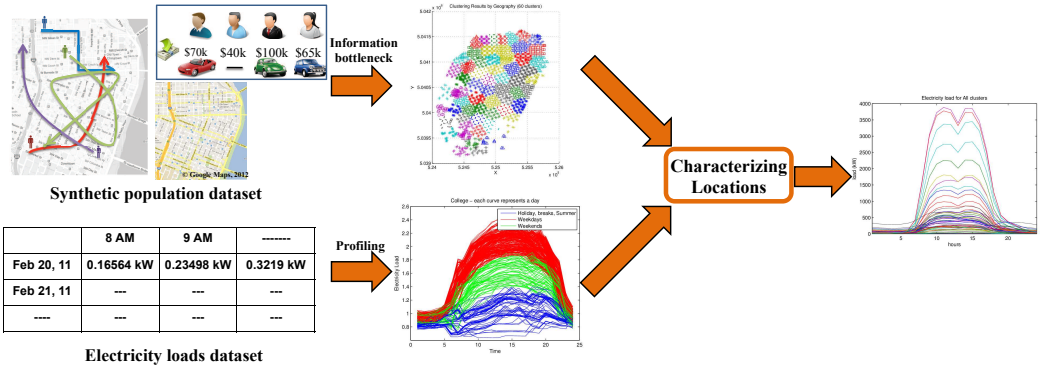
The synthetic dataset contains 243,423 locations of which 1,779 belong to the downtown area and of further interest for our purposes. Each location is represented by geographical [x,y] coordinate adopting the universal transverse mercator coordinate system (UTM) [Bisset et al. 2006]. There are a total 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 27 hour period (which includes a typical day) 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 travel 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).

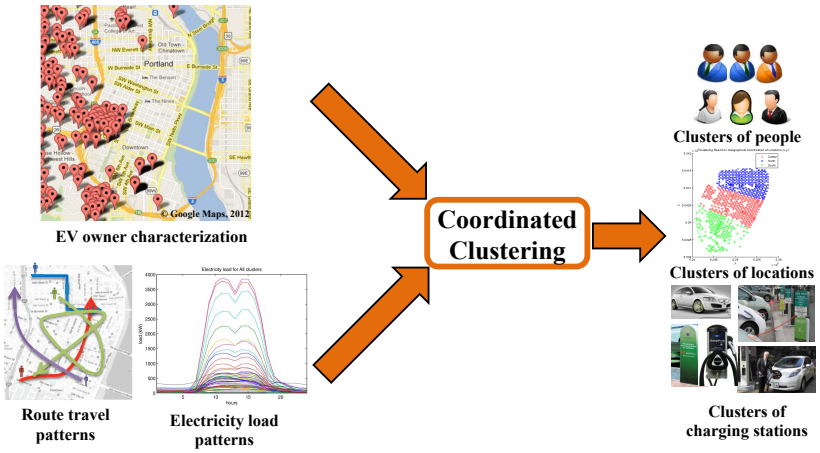
It is clear that the electricity load of each location greatly depends on the functionality of that location and hence our first approach is to utilize an information bottleneck type approach [Tishby et al. 1999] to characterize locations. Our aim is to cluster locations based on geographical proximity 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

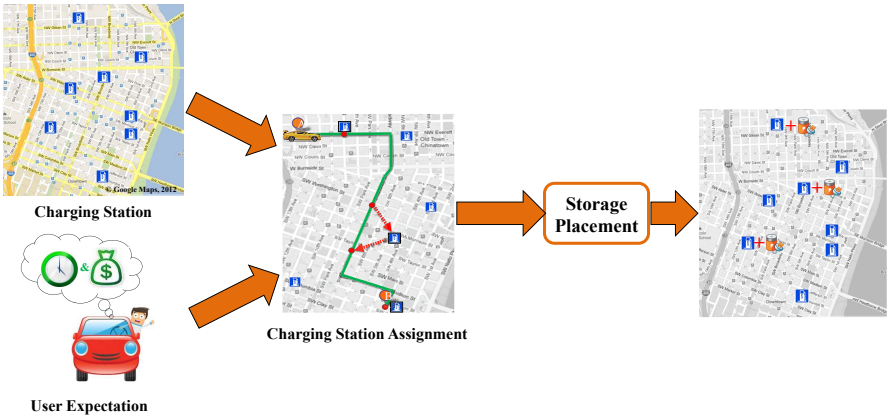


	8 AM	9 AM	-------
Feb 20, 11	0.16564 kW	0.23498 kW	0.3219 kW
Feb 21, 11	---	---	---
----	---	---	---

(a) Discovering location functionalities and characterizing electricity loads.



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



(c) Charging Station assignment and storage placement.

Fig. 1. Overview of our methodology.

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. In the proposed method, three datasets are used. Two datasets describe locations and one of

them describes people. Since each location has a set of features which do not depend on its coordination, we use one dataset to describe specific features of each location and another dataset that only consists of geographical coordinates. In addition to datasets that describe locations, we use a separate dataset for people with different income.

Choosing a right set of locations to install charging stations depends on many features. These features can be categorized into two groups: 1) Features of people who visit those locations. It is better to assume that these people have EVs, and because we assume that people with higher income have EVs, it is preferable to choose locations which people with higher incomes visit frequently. 2) Features of locations such as electricity load. In fact, we are looking for a set of candidate locations that have similar features and also, are visited by same type of people. Among different data mining approaches [Ramakrishnan and Grama 2001], clustering techniques can identify similarities and can categorize locations into different sets.

We use a coordinated clustering formulation to simultaneously cluster three datasets in a relational setting. Coordinated clustering tries to cluster different datasets such that relationships between items in each dataset are preserved. Here, we try to identify best locations to install charging stations where certain groups of people visit those locations. Candidate locations for charging stations are the ones that have specific characteristics such as low electricity load. However, we try to find those locations that have direct relationship with a specific group of people (people with high income). Obviously, type of features in people dataset is different from locations dataset. Due to this difference, and due to many-to-many relationship between locations and people we cannot use regular clustering approaches such as k-means. Our coordinated clustering framework builds upon our previous work [Hossain et al. 2010] 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 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.

In addition, we come up with the actual scenario for each EV owner, who needs charging to see which locations are the best ones for him with respect to charging cost and waiting time of EV owner. After measuring additional load of each charging station, we calculate the size of storage they need in order to reduce net load. Finally, we consider the economical aspects of storage deployment.

4. ALGORITHMS

As described in Section 3, our methodology comprises the following six major steps to determine candidate locations for charging stations:

i discovering locations' functionalities using an information bottleneck method;
ii electricity load estimation and integration with results of previous step;
iii studying the behavior of EV owners and calculating specific parameters relevant to their usage patterns;
iv candidate selection for charging stations using coordinated clustering techniques;
v finding appropriate charging stations for each user while maximizing user benefits; and
vi calculating the actual load of charging stations and storage placement.

Each of these steps are detailed next.

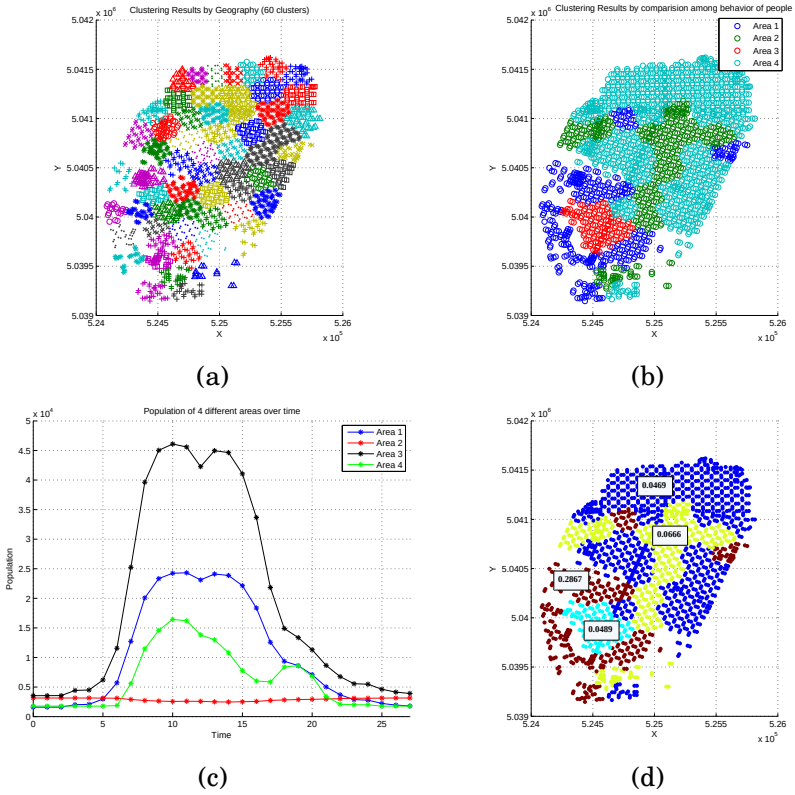


Fig. 2. (a) Clustering downtown locations based on geographic coordinates. (b) Clustering over the previous clustering with people's activities as side-information. (c) Dynamic population of the four discovered clusters over a typical day. (d) Computed residentiality ratio revealing one primary residential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

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. It is worth mentioning that in this part of our method, we desire to consider nearby locations and their electricity loads. Hence, the most appropriate approach for distance measurement is using Euclidean distance. 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

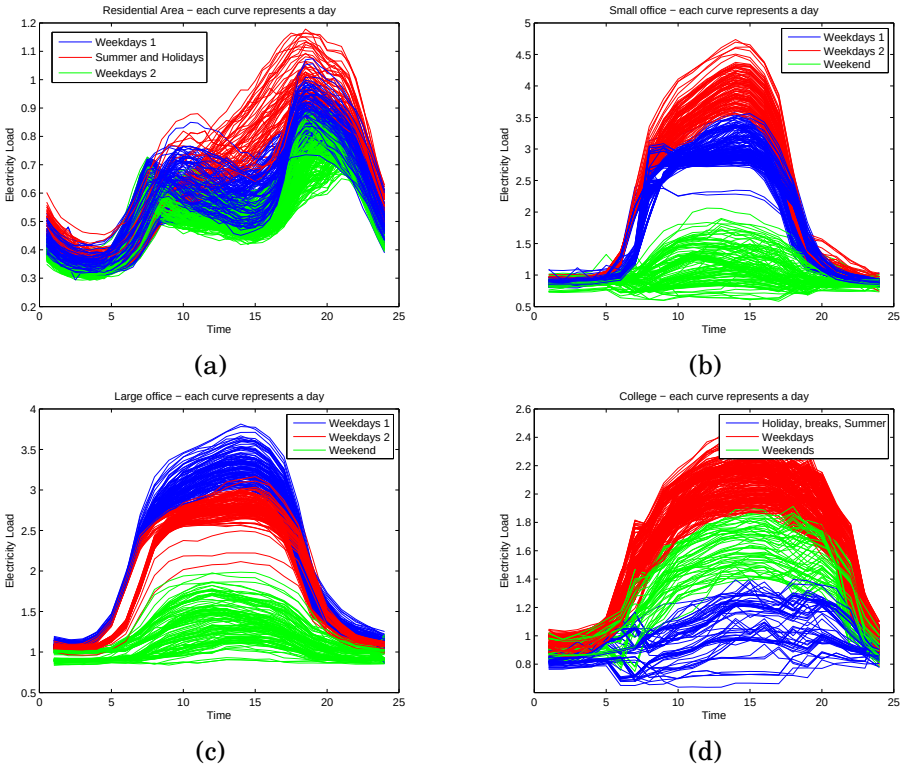


Fig. 3. (a) Electricity usage in residential areas. (b) Electricity usage in small office areas. (c) Electricity usage in large office areas. (d) Electricity usage in college areas.

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 using 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 [Blake et al. 2007], small office as 200sft [U.S. General Services Administration 1997], large office as 200sft [U.S. General Services Administration 1997], college as 50sft [The Engineering ToolBox], retail area as 50sft [The Engineering ToolBox], and

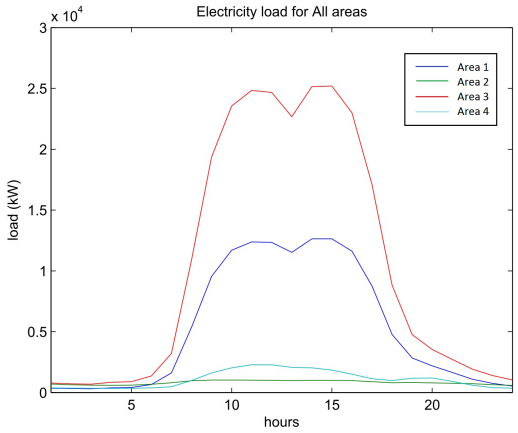


Fig. 4. Electricity loads for four characterized location clusters.

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 (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

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

where A_{p_i} is the average square footage per person for the purpose P_i and E_p is electricity consumption of building type p. It worth mentioning that E_p is from public data sources (California End User Survey (CEUS)) organized by NEC labs, America. 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 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 [Munro] and [Simply Hired, Inc.], 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 [KEMA, Inc. 2012] this is a realistic assumption if we consider different penetration scenarios in U.S in forecasted EV adoption in 2012-2022. Based on our model-

ing 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 for beginning/end of day situations, our goal is now to identify candidate charging locations during other times. Candidate charging stations will be a critical issue in near future as the number of EVs increases [Richard Martin 2012]. 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 [GM-Volt], 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 [GM-Volt], an electric car can go for 71.43 Km before it needs to be charged [The official U.S. Government Source for Fuel Economy Information].

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.

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

Charging needs is an estimation to see in which locations, EV owners will probably charge their EVs. It does not mean that vehicles will certainly charge at every location. Here, we say that "there is a greater chance for a location to be a charging station if people with higher charge needs visit that location". Equation (4) does not mean that vehicles will charge at every visit to such locations. It is just a measure that indicates which locations have higher chance of experiencing charging needs. For example, if location A is visited by 10 EVs where all of them need to be charged only once during a day, then charging need of location A would be 10. On the other hand, if location B is visited by 8 EVs in which all of them need to be charged twice, then the charging need of location B would be 16, which is higher than A's. These numbers show that with higher probability people in location B require charging needs, compared with A.

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.

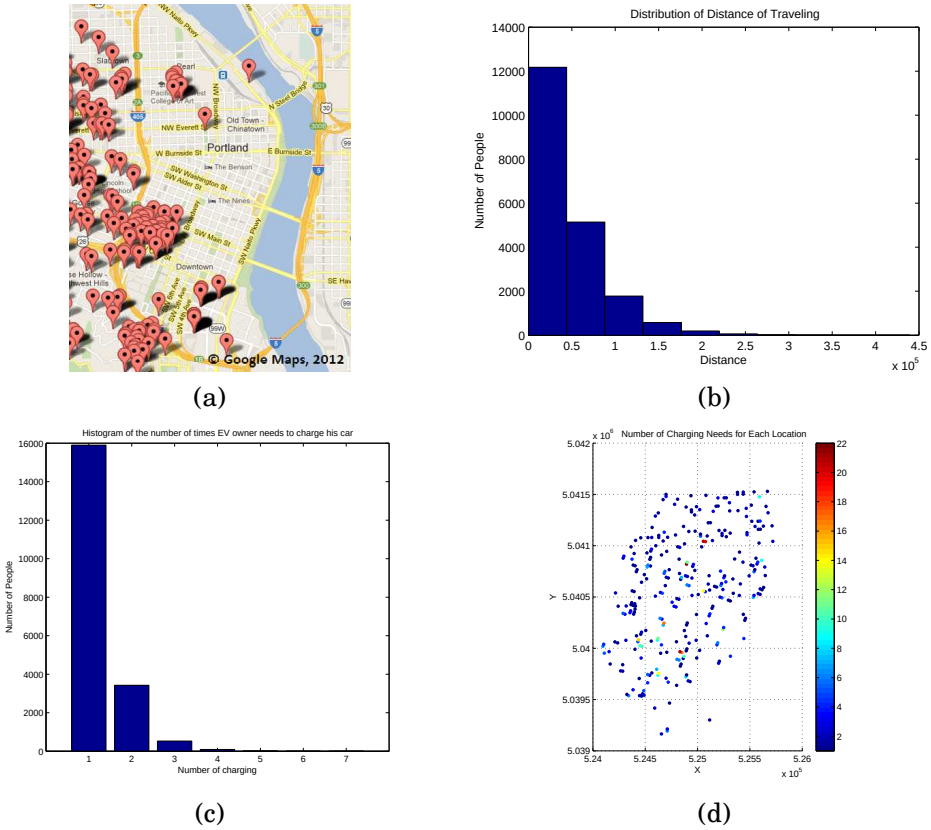


Fig. 5. (a) EV household locations. (b) Distribution of distances people travel in their EVs. (c) Charging needs for EVs. (d) Number of charging needs (more than zero) per location.

Table I. Table of Notations

Variable	Description	Variable	Description
B	relationship matrix	$C_{(x)}, C_{(y)}$	cluster indices
$\mathbf{m}_{i,\mathcal{X}}$, $\mathbf{m}_{j,\mathcal{Y}}$	prototype vectors for clusters	$v_i^{(\mathbf{x}_s)}$, $v_j^{(\mathbf{y}_t)}$	cluster membership indicators
k_x,k_y	number of clusters	w_{ij}	contingency table entry
$w_{i.}$	row-wise counts of contingency table entries	$w_{.j}$	column-wise counts of contingency table entries
α_i	row-wise random variable	β_j	column-wise random variable
U	uniform distribution	$\mathcal{F}$	objective function

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 [KEMA, Inc. 2012]. 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. In this section, we describe how coordinated

clustering can be used for charging station placement. Notations are summarized in Table I.

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, $\sum_{i=1}^{k_x} v_i^{(\mathbf{x}_s)} = \sum_{j=1}^{k_y} v_j^{(\mathbf{y}_t)} = 1$. 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 $v_j^{(\mathbf{y}_t)}$. Since our method operates over the prototypes, and uses membership probabilities to compute the probability distribution of the contingency table, it is mandatory that the functions are smooth and continuous everywhere in the system. These are the essential properties of our objective function. Any smooth and continuous membership function should work similarly. However, equation 5 has the advantage of involving Kreisselmeier-Steinhauser (KS) envelope function [Kreisselmeier and Steinhauser 1979] that is smooth and infinitely differentiable. As a result, our objective function can be optimized using any standard local and global optimizer.

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 probability distributions as follows

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

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

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:

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

where D_{KL} is the KL-divergence between two distributions and U indicates the uniform distribution over a row or a column. The idea of KL divergence is to estimate discrimination of information (Minimum Discrimination Information (MDI)) that leads us to use it as our divergence measure. Similar techniques that follow the MDI principle have the potential to be a part of our objective function. In the future, we plan to perform extensive experiments on this.

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.

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

where $p(\alpha)$ and $p(\beta)$ are defined as follows.

$$p(\alpha) = \frac{1}{n_x} \sum_{s=1}^{n_x} V^{(\mathbf{x}_s)} \tag{11}$$

$$p(\beta) = \frac{1}{n_y} \sum_{t=1}^{n_y} V^{(\mathbf{y}_t)}. \tag{12}$$

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.

$$\mathcal{F}_{\mathcal{X}\mathcal{Y}\mathcal{Z}} = \mathcal{F}_{\mathcal{R}}(\mathcal{X}, \mathcal{Y}) + \mathcal{F}_{\mathcal{R}}(\mathcal{Y}, \mathcal{Z}). \tag{13}$$

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

4.5. Charging Station Assignment based on User Expectations

After determining candidate charging stations, we need to assess the effect of installing charging stations at those locations, and evaluate the changes in electricity load. In addition, from a business point of view, it is important to study the size of storage needed at those locations.

First, we need to evaluate candidate charging stations resulting from our co-clustering algorithm. One solution is to see whether these set of candidates are even used by EV owners. In order to understand which locations tend to be charging stations from EV owners's point of view, we need to identify the desired locations of each person. These locations are the ones that minimize cost of charging for EV owners. On the other hand, since the process of charging an EV typically will take several hours, user tends to charge his car in those locations where he stays for at least a few hours. Obviously, from a business point of view, we not only consider those locations that meet users' criteria (charging cost), but also aim to optimize charging station in terms of electricity load and size of storage.

In this subsection, we show how to determine where users desire to charge their vehicles with respect to cost of charging and change of route for each user. In what follows, we develop an algorithm to assign charging stations to users. Of course, users have the freedom to select their charging stations. We assume that they are intuitively looking for the cheapest options. Also, we assume that users desire to minimize their detour and their waiting time (for charging). These assumptions were considered in the assignment algorithm. This assignment requires an estimate of storage sizes of charging stations. For this reason, we need to know the exact schedule of users to calculate the overall electricity load of each location. We assume that detour, cost, and waiting time are important issues in selecting charging stations for all users. (It should be noted that the goal here is to estimate the storage size, not to suggest charging stations to users.)

ALGORITHM 1: User-based Candidate Charging Stations (UCCS)

Input: $Route$ consists of sequence of locations.
Output: ChargingStations consists of best locations to charge as well as level of charging and time of charging at those locations.
CS= $RUCCS(Route(1), R, Route)$ `/* assume at first each car has fully charged (R) */`
$MinFailure = \min(\text{CS}(Failure))$;
$MinFailureSet =$subset of CS with $Failure$ equal to $MinFailure$;
$ChargingStations = arg\min(MinFailureSet(Cost))$;
return $ChargingStations$

To the best of our knowledge, there is limited work on the "where to charge" problem in the literature. In [Khuller et al. 2011], authors try to find the cheapest tour between customer destination locations to fill gas. Our work is different from [Khuller et al. 2011] for a variety of reasons. For example, in our problem,

(1) Sequence of stop points for each user is determined.
(2) We do not have a boundary on the number of times that an EV owner can charge his car.
(3) Price of charging in each location varies based on duration of stay of user in that location.
(4) In some locations, car battery will be charged partially.

Before explaining our algorithm, it is worth mentioning that there are different standards for charging stations. Charging time of each EV depends on its capacity and the charging level of the charger. Levels of charging for EVs can be categorized into three levels: level 1, level 2, and level 3 (DC power). Power consumption of each level is different from each other and hence, prices are different. Furthermore, rate of charging (the time that it takes to charge a battery for 1 KWh) is different for each level.

The algorithm for estimating desired charging locations based on user point of view is as follows:

For each user, we invoke Algorithm 1 (UCCS). This algorithm takes the route of one user as input and calculates the best locations for charging as well as level of charging and the time of charging. Algorithm 1 calls Algorithm 2 to compute all feasible sets of charging stations in the route that user travels. Then, Algorithm 1 only retains those sets that have minimum number of failures, i.e. minimum number of times that car has to switch to gas because of empty battery. After that, it selects a set of charging stations which has a minimum cost of charging.

Algorithm 2 (RUCCS) is a recursive function for finding all feasible sets of charging stations. It takes the current location, remaining charge in the EV, and the route of user as inputs and calculates sets of candidate charging stations. This algorithm works as follows:

Let us assume that currently the EV is at location L_j, and that the available charge of battery is equal to C_j. Also, assume that d, the distance that the EV can travel from L_j without charging its battery, can be computed. This distance is determined in Line 2. Here, R is the capacity of the battery and D is the distance that EV can travel with a fully charged battery. In Line 3 of the algorithm, we determine A as the set of locations that are located on the route of EV, and are at most d meters away from L_j. It is obvious that if the last point of the route is in A, we do not need to re-charge the battery (Lines 4-6). On the other hand, the EV must re-charge its battery in at least one of the locations in A; otherwise after d meters, it should switch to gas.

However, when A is empty, there is no way to re-charge the battery of EV. In that case, EV must switch to gas and we say that a $failure$ has happened. After a failure,

in the next subsequent stop point, L_{j+1}, EV's battery must be re-charged. In this case, we recursively call RUCCS for L_{j+1} (Lines 8-23). Here, $MaxC_{j+1,k}$ is the maximum possible charge of battery which is determined based on duration of stay of the car, and level of charge, k. However, because the capacity of battery is R, the actual value of the charge is calculated in line 11 and is shown by $C_{j+1,k}$. Cost of this charge is determined in line 12 by $CostC_{j+1,k}$.

If A is not empty (line 24), we must choose the most feasible location in A for re-charging the battery. Therefore, in Lines 25-32, for each location in A, and for each charging level, k, we calculate the amount of possible charge in that location ($C_{i,k}$), the cost of charging ($CostC_{i,k}$), and the maximum distance that the car will travel ($MaxD_{i,k}$), if we charge it in that location with that charging level. For each charging level, the best stop point for re-charging the car is the one that if we re-charge our vehicle there, we can travel further with respect to the current location, L_j.

Choosing the best members of A for re-charging is performed in Line 33. Then, if the best stop point for charging level k is L_{idx}, we recursively call RUCCS with inputs L_{idx}, $C_{idx,k}$, and $Route$ (Line 34). After returning from a recursive call of RUCCS for a location such as L_i, (Lines 13 and 34), we have several sets of stop points that are considered as feasible sets located after L_i. These sets are determined with this assumption that L_i is a charging station too. Hence, we have to add L_i to all of these sets before returning from the current iteration of the algorithm (Lines 14-19 and 35-40). Also, because we want to consider all feasible solutions to choose the best one, we have to keep all the results that are determined for different charging levels. This step is performed in Lines 20 and 41.

After determining the most feasible locations from the users perspective, i.e. locations that minimize charging cost and number of failure's, we must match existing charging stations with the new locations. Since, we cannot establish charging station for each location that users want, we choose those charging stations that were extracted from Section 4.4 and assign each user to them based on distance to charging stations. Hence, for each charging station, we know when and how many times it will serve EVs. In order to select the best charging stations for a user, we use a *nearest charging station* assignment policy. Therefore, if the desired location for charging is L_i and S_c is the set of available charging stations we use C_i instead of L_i where

$$C_i = arg\min distance(L_i, C_j) \text{ for all } C_j \text{ in } S_c \tag{14}$$

where, $distance(A, B)$ measures the distance between locations A and B. It should be mentioned that any method of distance measurement (Euclidean, Manhattan, ...) can be used in this function.

With this policy, detours are minimized. After assigning charging stations, the amount of electricity load added to charging stations based on their serving time will be calculated.

ALGORITHM 2: Recursive Function (RUCCS)

Input: L_j is the current location, and C_j is available charge of car at location L_j and $Route$ consists of sequence of locations.

Output: CS which consists of sets of candidate charging stations. Each candidate charging set (CS_i) has the following fields:
$CS_i(points)$ is the ordered set of locations where user must charge his car.
$CS_i(level)$ is level of charging at each location in $CS_i(points)$.
$CS_i(costs)$ is cost of charging at each location.
$CS_i(Failure)$ is the number of failure during trip.

CS $= \{\}$;
$d = C_j * \frac{D}{R}$
$A =$ set of stop points in distance d of L_j;
if *Route(end) is in* A **then**
  return CS;
end
if $|A| = 0$ **then** `/* failure will happen and it must switch to gas */`
  $L_{j+1} =$ next subsequent stop point in Route;
  for $k = 1$ **to** 3 **do**
    $MaxC_{j+1,k} =$maximum possible charge at L_{j+1} with level k;
    $C_{j+1,k} = \min(MaxC_{j+1,k}, R)$;
    $CostC_{j+1,k} =$cost of charging at L_{j+1} with level k;
    $CS^k = RUCCS(L_{j+1}, C_{j+1,k}, Route)$;
    for *each candidate set,* CS^k_m*, in* CS^k **do**
      $CS^k_m(points) = [L_{j+1} \quad CS^k_m(points)]$;
      $CS^k_m(levels) = [k \quad CS^k_m(levels)]$;
      $CS^k_m(costs) = [CostC_{j+1,k} \quad CS^k_m(costs)]$;
      $CS^k_m(failure) = CS^k_m(failure) + 1$;
    end
    CS=CS $\cup$ CS^k;
  end
  return CS;
else
  for *each point,* L_i*, in* A **do**
    for $k = 1$ **to** 3 **do**
      $MaxC_{i,k} =$maximum possible charge at L_i with level k;
      $C_{i,k} = min(C_j - \frac{dist(L_j,L_i)*R}{D} + MaxC_{i,k}, R)$;
      $CostC_{i,k} =$cost of charging at L_i with level k;
      $MaxD_{i,k} = \frac{D}{R} * C_{i,k} + dist(L_j, L_i)$;
    end
  end
  for $k = 1$ **to** 3 **do**
    $idx = arg\max_{L_i \in A}(MaxD_{i,k})$;
    $CS^k = RUCCS(L_{idx,k}, C_{idx,k}, Route)$;
    for *each candidate set,* CS^k_m*, in* CS^k **do**
      $CS^k_m(points) = [L_{idx,k} \quad CS^k_m(points)]$;
      $CS^k_m(levels) = [k \quad CS^k_m(levels)]$;
      $CS^k_m(costs) = [CostC_{idx,k} \quad CS^k_m(costs)]$;
      $CS^k_m(failure) = CS^k_m(failure) + 1$;
    end
    CS=CS $\cup$ CS^k;
  end
  return CS;
end

4.6. Storage Placement

In previous section, we determined profile of electricity load at each location before and after charging station deployment. Profile of electricity load after installing charging stations is determined based on number of cars that are charged at each location and their corresponding level and duration of charging. On the other hand, each location has a predetermined capacity which is the maximum electricity load that it can tolerate. When electricity load of a location increases and goes above its capacity, we need to place storage to meet the electricity demand of that location. In this regard, the efficiency of storage is also important. Here, we assume that the desired utilization of storage in all locations is 80% i.e. at most 80% of the capacity of a storage is used in a day. That ensures us that storage will not discharged to no more than 80% of total capacity. Due to the small size of storage at some locations we aggregate storages of nearby locations. For this purpose, we use DBSCAN [Ester et al. 1996] to locate dense areas and calculate the needed storage size of each cluster as a summation of storages over all locations in that cluster.

From a business point of view, placing storage at a charging station must have a adequate revenue for storage owners. In addition, putting storage at locations is advantageous to city in terms of reducing the peak of electricity load in urban area.

To investigate the revenue of storage units, we consider each charging station in turn and compute the revenue of storage. Here, revenue refer to the amount of funds that storage owners will save from selling energy to consumers. Revenue can be achieved by selling energy during the day and recharging the storage during the night (with off-peak rate). In addition, to observe profile of charging stations based on their load curves, we use the clustering algorithm introduced in [Yang and Leskovec 2011], i.e., the K-Spectral Centroid (K-SC) algorithm for time series data using a similarity metric invariant to scaling and shifting. They apply adaptive wavelet-based incremental approach to K-SC to use it for large datasets. K-SC proved to be an effective clustering when scaling is not important. By applying this method, we can understand different types of charging stations based on their load curves and finally, locate the best locations to put storage in order to get high revenue.

5. RESULTS

Figure 6 describes the coordinated clustering scenario. As illustrated in this figure, we use three datasets: People (Income), coordinates (x,y) of location, and features (load, charge need, stay) of location. First dataset contains information about income of people and second dataset has information about the geographic coordinates of each location. However, the third dataset contains the characteristics of locations for charging station placement. Electricity load of buildings, charging need of people in that location and duration of stay in each location are three features in this dataset.

We begin with some preliminary observations about our data. Figure 7 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 8 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 8 (b).

Profiles of these clusters are described in detail in Figure 9. Of particular interest to us is the view from the perspective of EV attributes, i.e., Figure 9 (c). Details of these clusters are explored in greater detail in Table III. 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),

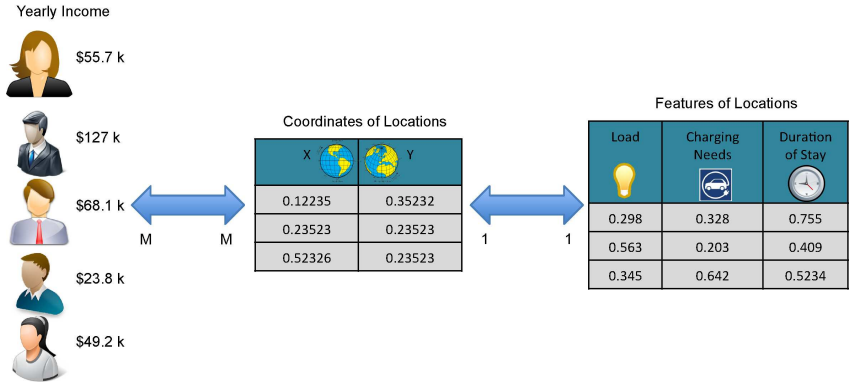


Fig. 6. Coordinated clustering schema.

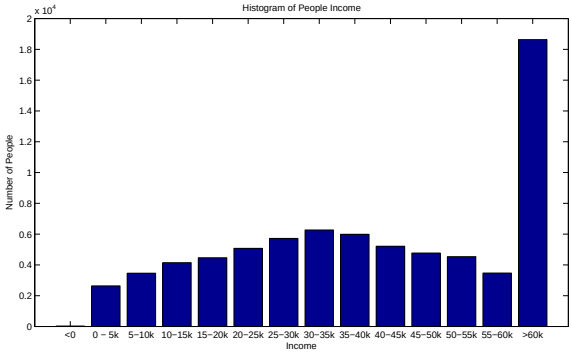
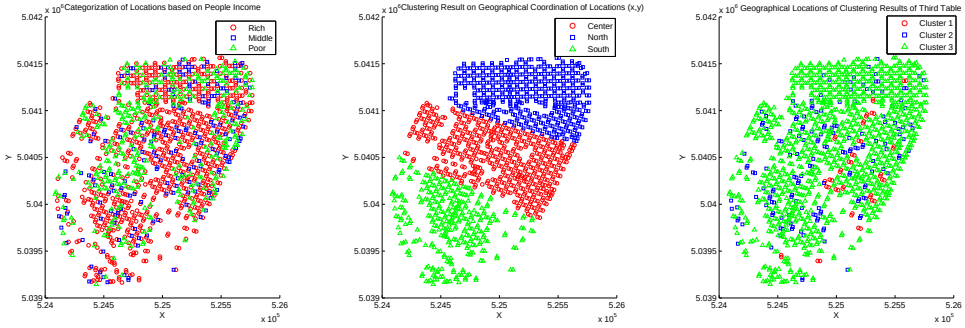


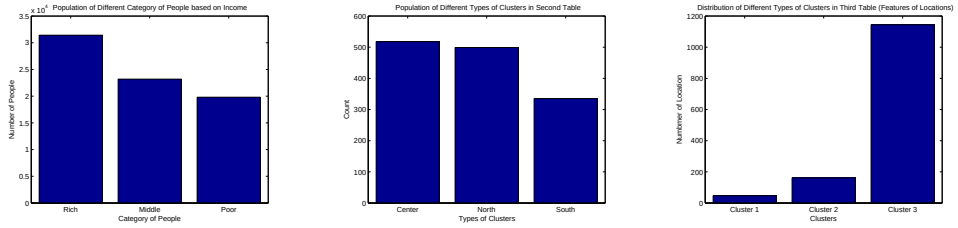
Fig. 7. Distribution of income.



(a) (b) (c)

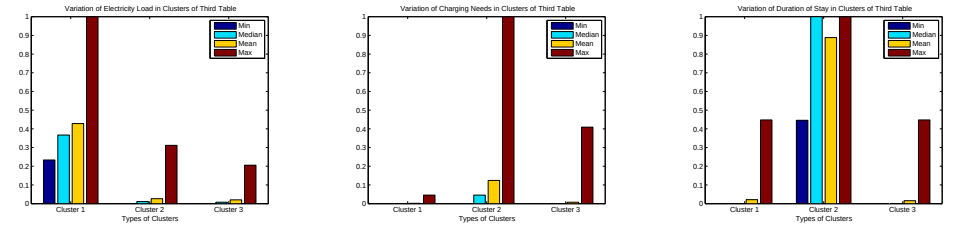
Fig. 8. Results of coordinated clustering (3 clusters) when viewed through the attributes of each domain. (a) Clusters based on income. (b) Clusters based on geographical location. (c) Clusters based on EV charging station attributes.

and high staying duration [KEMA, Inc. 2012]. As can be seen from Table III cluster 2 from the third dataset fits these requirements. Greater insights into the three clusters from the viewpoint of these three attributes is shown in Figure 10 supporting the choice of locations in cluster 2 as the right candidates for locating charging stations. As we mentioned before, we try to identify locations with specific features while certain group of people (people with high income) visit those locations. Although based on the clusters of first dataset (people), we must choose locations where mostly people with



(a) (b) (c)

Fig. 9. 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)

Fig. 10. 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.

Table II. Profiles of Clusters in Third Dataset (Location's Features)

Cluster	% of People with High income	% of Locations with High Elec. load	% of Locations with High Charging need	% of Locations with High Stay
1	0.43	0.45	0	0.02
2	0.41	0.06	0.15	0.88
3	0.41	0.05	0.01	0.01

greater salary affordances visit, the distribution of high income vs. low income people in clusters 1, 2 and 3 in third dataset (locations) are almost similar. This is illustrated in Table II. With respect to distribution of high income people people, cluster 1 is better selected. However, cluster one is not a good choice for installing charging stations because 45% of its locations are those with high electricity load. Between other two clusters (cluster 2 and cluster 3), cluster 2 is better because it has low electricity load, high charging need, and high duration of stay.

With the aid of clustering, we can predict which locations are the best candidates to install charging stations. However, the effect of installing charging stations in these locations on other metrics such as the price of charging and electricity load of buildings must be evaluated.

Since we are looking only at downtown area of Portland, we do not have any information about exact location of other charging stations outside of downtown. Here, we padded our downtown area by 500 meters from each side (if we suppose downtown has a rectangular shape). Then, assuming that those cars in the padded area can be served by our current charging stations, we run the algorithm 1 for each car. The distance between charging station and current location of car must be minimized because charging at charging station with Level 1 or 2 will take several hours and people prefer to charge their cars at those locations that they stay longer. In reality, users

Table III. Characteristics of Clusters in Third Dataset (Location's Features)

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

Table IV. Characteristics of Charging Stations

Level	Description	Elec. Load(kW)	Cost(¢/kWh)			Time(h)
			on-peak	mid-peak	off-peak	
1	110v outlet,16 Amp	2.2	16.62	10.85	7.77	8
2	220v charger, 16 Amp	3.3	16.62	10.85	7.77	4
3	400v DC, 125 Amp	50	10.89	6.36	3.63	0.5

can charge their cars anywhere in vicinity ($\sim$1 mile) of their desired buildings (ex. he can park his car at nearest charging station and walk to his office). Furthermore, we need to have information for two types of movements (riding to charging station, and walking to office). Since, the distances are not too large, using Euclidean distance to measure distances is not troublesome and makes computations easier. Furthermore, the actual information about roads of the area was not available to use and the dataset consists only origin and destination of each movement.

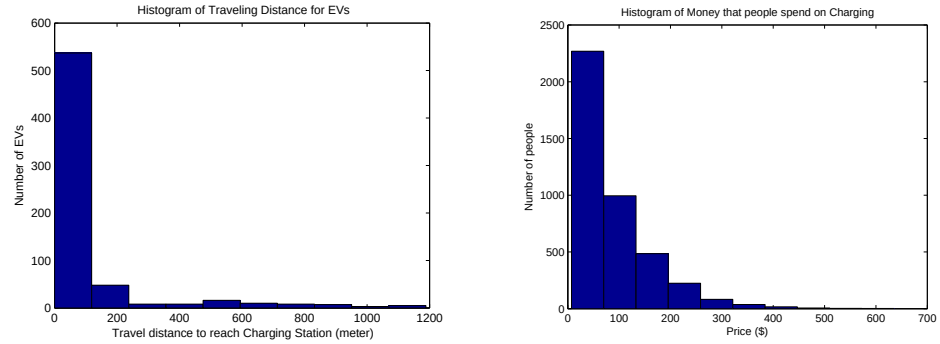
Specifications of three levels of charging for Portland are summarized in Table IV based on PGE [Portland General Electric Company 2012a] and [Portland General Electric Company 2012b]. From [Portland General Electric Company 2012b], Schedule 7 is chosen for level 1 and 2 and Schedule 32 is selected for level 3. It is worth mentioning that prices (tariff rate) are based on time of use policy (TOU). The definition of On-peak, Mid-peak, and Off-peak is inspired from the electricity loads in our dataset:

— On-peak: 6 AM to 10 AM and 5 PM to 8 PM
— Mid-peak: 10 AM to 5 PM and 8 PM to 10 PM
— Off-peak: 10 PM to 6 AM

Prices at Table IV are for both buying electricity from the grid and from charging station (by EVs). The type of charging depends on time of stay. If an EV stays for 8 hours, it can charge by charging level 1 which is cheapest option. If an EV stays for 4 hours, it can use level 2 charger whereas if the EV needs to be charged in 30 minutes, it can use the level 3 (DC) option. Price of charging in level 3 is very high compared to level one and two. For example, cost difference (a complete charging) between charging by DC and level I would be $50*10.89-2.2*16.62 = 507.936$ cents or \$5. Hence, the overall impact of level of charging (I, II, and DC) is very high on charging stations and on users in cost and electricity load points of view.

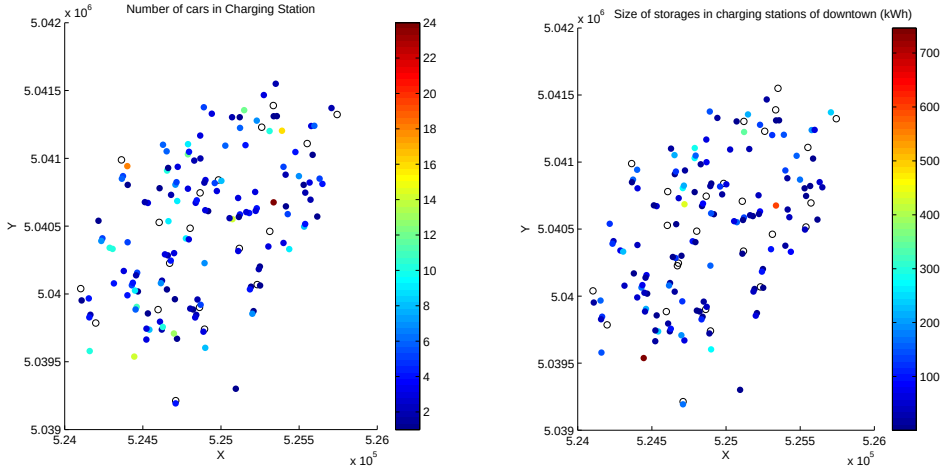
Experiments show that average distance traveled by each car is 8.4881 meters and that the maximum distance traveled in this experiment was 1188.2 meters. Figure 11 (a) depicts the histogram of distance between location of current stop point and available charging station. This result is promising since we considered part of the boundaries of downtown while there might be a charging station in that area. The number of charging stations based on our clustering algorithm is 161 while number of locations that people liked to charge their cars is 367. The histogram of expenses that all EV owners in Portland will pay daily for charging is shown in Figure 11 (b).

Also, number of cars that are served at each charging station is important from a business point of view, to study revenue of charging station owners. As Figure 12 (a) shows this is zero for some charging stations (black circles) and they can be removed from consideration as charging station candidates. Based on this figure, we can place



(a) (b)

Fig. 11. (a) Histogram of distance between current stop point of location and available charging station (meter). (b) Histogram of expenses people spend on charging during a day.



(a) (b)

Fig. 12. (a) Number of cars that served by each charging station. Note that some charging station (black circle) are useless and can be removed. (b) Size of storage in charging stations (kWh). Note that some charging station (black circle) are useless and can be removed.

appropriate charging infrastructure at those locations that serve certain number of cars.

It should be noted that the number of failures in our algorithm is 48. This highlights the number of cases where an EV must switch to gas in order to continue its route. Experiments show that all of these 48 cases was due to the nature of our dataset, i.e. distance between locations was more than maximum possible distance of travel with a full battery.

Those charging stations that provide service to cars will add extra load to the location. This load might be more than the capacity of the location. Here, we assume that maximum load of one location during a day is equal to its capacity. We must place storage to those locations that require extra electricity. For determining the size of storage at each charging station, we must look at values of location's capacity and electricity load after adding EV. To compute size of storages, we would like to assume that stor-

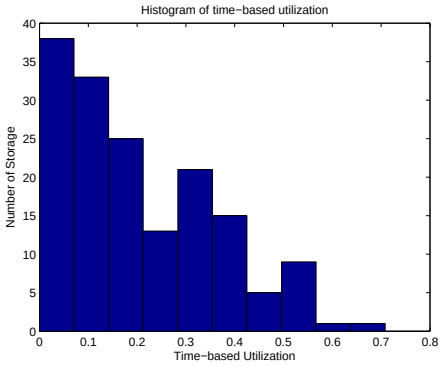


Fig. 13. Histogram of Time-based Utilization.

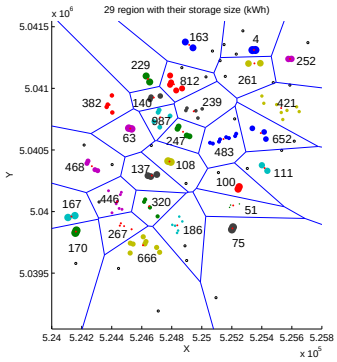


Fig. 14. Aggregated regions. Value of storage size (per kWh) is shown for each region (sum of all storages)

age will not discharge to more than 80%. For example, 50kWh should be selected for a net daily load of 40kWh. The size of required storage should be calculated from the area below the curve of new electricity load (kW X hr) and above the capacity (net peak load)(kW). Typically, storage will be charged at night and used during the day and it should be sized to cover a day's net load.

Figure 12 (b) shows how many locations need to have storage. Again, black circle means there is no need for storage at this location. Based on our assumption, utilization is 80%. However, the time-based utilization (i.e. the percentage of time that storage has been used in a day) is shown in Figure 13. Obviously, the value of time-based utilization cannot be 1 in this case, because storage needs to be recharged over night to be used for the next day.

After determining storage sizes, we can aggregate them to minimize number of storage units. This aggregation is based on vicinity of locations. Hence, we used DBSCAN to find dense areas and take summation of the storage size over all locations in each cluster. This is shown in Figure 14. In this figure, small black circles represent those locations where they couldn't be grouped by other locations and considered as noise in DBSCAN algorithm. small red dots represent center of each group. As an example, in upper-right side, overal storage size of two locations (purple circles) is 252 kWh.

To study the amount of saving for each storage, we assume that each storage will charge at night with off-peak rate. Hence, during the day in on-peak and mid-peak hours, storage owners will sell energy to consumers (EV owners). Hence, the difference between price of selling and price of recharging will be considered as revenue of

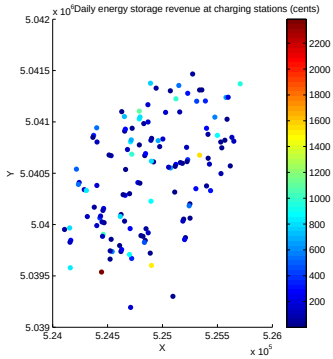


Fig. 15. Revenue of energy storage at charging stations

storage. Figure 15 shows energy storage revenue for each charging station that has storage unit. This value of revenue is calculated for one typical day.

To observe the profiles of charging stations based on their load curves after adding EVs and after adding storage, we used the K-SC clustering approach described earlier. Here, the value of electricity load before adding EV, after adding EV, and after storage deployment during 24 hours were considered as a vector of 24*3 elements. The profiles of charging stations are categorized into 4 clusters which the prototype of each cluster is shown in Figure 16. It should be noted that this clustering is invariant to shift and scale and that is why the value of load after storage deployment is higher than maximum value of load before considering EV. Figure 17 depicts an example of actual curves for one charging station in cluster 1. It is obvious that storage deployment will ensure that the value of electricity load will not go higher than the capacity at each location. Figure 16 is important in understanding the behavior of charging stations. Also this figure is helpful in deciding between using a mobile storage unit and a stationary one.

In Figure 16, charging stations in clusters 1 and 4 have little impact on the peak load, whereas those in cluster 2 and 3 significantly increase peak demand of the system. Therefore, using energy storage for charging stations in cluster 2 and 3 would make more sense than in clusters 1 and 4. Based on number of charging stations in each cluster, 43% of charging stations (in cluster 2 and 3) are candidates for storage deployment. On the other hand, if there is no possibility of adding energy storage, charging stations in clusters 1 and 4 would have much less impact on the grid and will be accepted by utilities with less opposition. Also, one can deploy mobile storage units for charging stations in clusters 1 and 4.

Figure 18 shows the amount of daily revenue achieved by storage deployment for each cluster. In this figure, locations in cluster 2 and 3 have highest revenue compared with cluster 1 and 4. Total revenue in cluster 1 and 4 is 6547.2 cents while total revenue in cluster 2 and 3 is 33081.0 cents. Based on this, one can justify using stationary battery storage in candidate charging stations (cluster 2 and 3).

6. DISCUSSION

Electrical vehicles are 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 well as storage infrastructure as EV needs grow. In addition, we identified candidate locations for deployment of stationary energy storages to utilize existing electricity infrastructure. The results presented here can be generalized to a temporal scenario where we accommodate a growing EV pop-

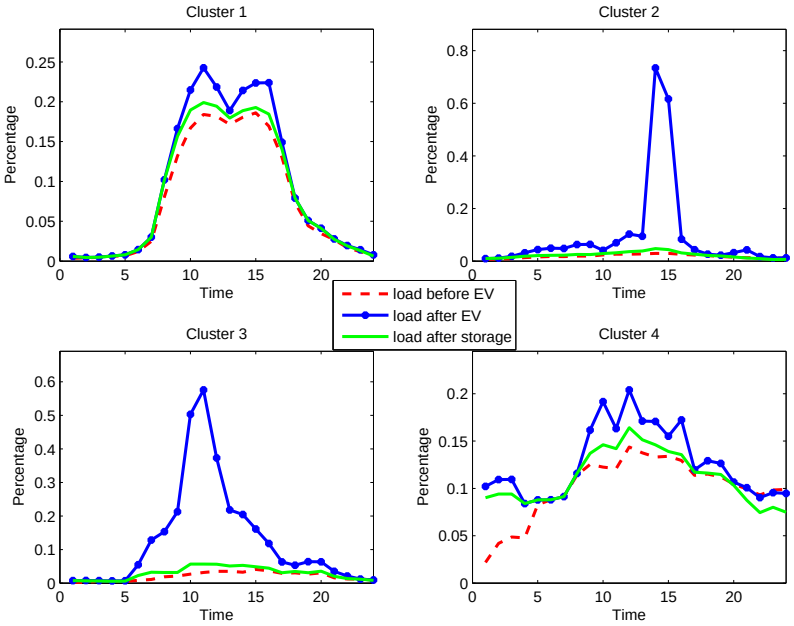


Fig. 16. Profiles of four types of charging stations based on load curves

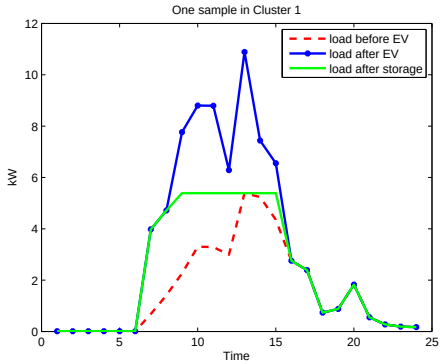


Fig. 17. Profile of one charging station in cluster 3

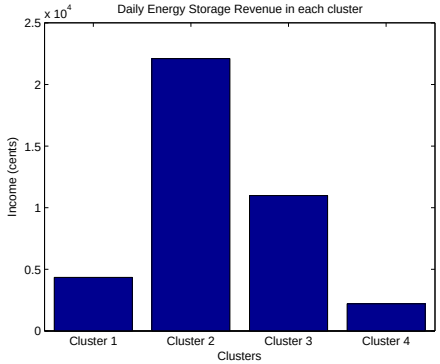


Fig. 18. Revenue of energy storage at each cluster

ulation and to design charging infrastructure to accommodate additional scenarios of smart grid usage and design.

The methodology presented in this paper mostly incorporates demand data from the electricity infrastructure and future work would incorporate information from the electricity supply side too. Information such as loading level of electricity feeders and remaining excess capacity of feeders for EV charging stations can be integrated in the methodology to improve the placement of EV charging stations. Also, there are several measures that were not considered here, such as life of battery, peak shaving reduction, adding PVs to current system, and details of economic analysis in charging stations and energy storage deployment. Incorporating these aspects is a direction of future work. Finally, the analysis presented here integrates a small range of datasets, each of which has adequate coverage over regions of interest. To overcome regions of data sparsity, we could employ the use of surrogate models like Gaussian processes [Ramakrishnan et al. 2005], which can enable the integration of a greater variety of datasets.

REFERENCES

AMAN, S., SIMMHAN, Y., AND PRASANNA, V. K. 2011. Improving Energy Use Forecast for Campus Microgrids using Indirect Indicators. In *IEEE Workshop on Domain Driven Data Mining*.

BAILEY-KELLOGG, C., RAMAKRISHNAN, N., AND MARATHE, M. 2006. Spatial data mining to support pandemic preparedness. *SIGKDD Explor. Newsl. 8,* 1, 80–82.

BAYRAM, I., MICHAILIDIS, G., DEVETSIKIOTIS, M., BHATTACHARYA, S., CHAKRABORTTY, A., AND GRANELLI, F. 2011. Local Energy Storage Sizing in Plug-in Hybrid Electric Vehicle Charging Stations under Blocking Probability Constraints. In *IEEE International Conference on Smart Grid Communications (SmartGridComm)*. 78–83.

BISSET, K., ATKINS, K., BARRETT, C., BECKMAN, R., EUBANK, S., MARATHE, A., MARATHE, M., MORTVEIT, H., STRETZ, P., AND KUMAR, V. A. 2006. Synthetic Data Products for Societal Infrastructures and Proto-Populations: Data Set 1.0. Tech. Rep. TR-06-006, Network Dynamics and Simulation Science Laboratory, Virginia Tech, Blacksburg, VA.

BLAKE, K. S., KELLERSON, R. L., AND SIMIC, A. 2007. Measuring Overcrowding in Housing. U.S. Department of Housing and Urban Development, Last accessed Oct 13 2012 `http://www.huduser.org/publications/polleg/overcrowding_hsg.html`.

ESTER, M., KRIEGEL, H., SANDER, J., AND XU, X. 1996. A Density-based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *KDD '96*. 226–231.

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

GUISEPPE, M. AND ANTONIO, V. 2012. Fast charging stations for electric vehicle: The impact on the mv distribution grids of the Milan metropolitan area. In *2012 IEEE Energy Conf. Exhibit.* 1055–1059.

HOFFMAN, M., SADOVSKY, A., KINTNER-MEYER, M., AND DESTEESE, J. 2010. Analysis Tools for Sizing and Placement of Energy Storage in Grid Applications. Tech. Rep. PNNL-19703, Pacific Northwest National Laboratory Richland, Washington 99352 .

HOSSAIN, M. S., TADEPALLI, S., WATSON, L. T., DAVIDSON, I., HELM, R. F., AND RAMAKRISHNAN, N. 2010. Unifying Dependent Clustering and Disparate Clustering for Non-homogeneous Data. In *KDD '10*. 593–602.

KEMA, INC. 2012. Distributed Energy Storage: Serving National Interests, Advancing Wide-Scale DES in the United States. Tech. Rep. 20130065, National Alliance for Advanced Technology Batteries.

KHULLER, S., MALEKIAN, A., AND MESTRE, J. 2011. To fill or not to fill: The gas station problem. *ACM Transactions on Algorithms (TALG) 7,* 36, 1–15.

KINDBERG, T., CHALMERS, M., AND PAULOS, E. 2007. Urban Computing. *IEEE Pervasive Computing*, 18–20.

KREISSELMEIER, G. AND STEINHAUSER, R. 1979. Systematic Control Design by Optimizing a Vector Performance Index. In *IFAC Symp. on Computer Aided Design of Control Systems*. 113–117.

LIU, W., ZHENG, Y., CHAWLA, S., YUAN, J., AND XIE, X. 2011. Discovering Spatio-Temporal Causal Interactions in Traffic Data Streams. In *KDD '11*.

MAKAROV, Y., PENGWEI, D., KINTNER-MEYER, M., CHUNLIAN, J., AND ILLIAN, H. 2012. Sizing Energy Storage to Accommodate High Penetration of Variable Energy Resources. *IEEE Transactions on Sustainable Energy 3,* 1, 34–40.

MOMTAZPOUR, M., BUTLER, P., HOSSAIN, M. S., BOZCHALUI, M. C., RAMAKRISHNAN, N., AND SHARMA, R. 2012. Coordinated clustering algorithms to support charging infrastructure design for electric vehicles. In *Proceedings of the ACM SIGKDD International Workshop on Urban Computing*. UrbComp '12. ACM, New York, NY, USA, 126–133.

MUNRO, N. 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/.

PAUL, T. AND AISU, H. 2012. Management of Quick Charging of Electric Vehicles Using Power from Grid and Storage Batteries. In *2012 IEEE Intl. Electric Vehicle Conf.* 1–8.

PORTLAND GENERAL ELECTRIC COMPANY. 2012a. Charging on the Go. Last accessed Oct 13 2012 http://www.portlandgeneral.com/community_environment/initiatives/electric_vehicles/charging_stations/charging_on_go.aspx.

PORTLAND GENERAL ELECTRIC COMPANY. 2012b. PGE's retail Rate Schedules and Rules. Last accessed Sep 21 2012 www.portlandgeneral.com/our_company/corporate_info/regulatory_documents/tariff.

RAMAKRISHNAN, N., BAILEY-KELLOGG, C., TADEPALLI, S., AND PANDEY, V. 2005. Gaussian processes for active data mining of spatial aggregates. In *Proceedings of the SIAM International Conference on Data Mining*.

RAMAKRISHNAN, N. AND GRAMA, A. 2001. Mining Scientific Data. *Advances in Computers Vol. 55*, 119–169.

RAMCHURN, S. D., VYTELINGUM, P., ROGERS, A., AND JENNINGS, N. R. 2012. Putting the 'Smarts' into the Smart Grid: A Grand Challenge for Artificial Intelligence. *Communications of the ACM 55,* 4, 86–97.

RICHARD MARTIN. 2012. More than 11 Million EV Charging Stations Will be Installed Worldwide by 2020. Last accessed Oct 13 2012 http://www.pikeresearch.com/newsroom/more-than-11-million-ev-charging-stations-will-be-installed-worldwide-by-2020.

ROSSWOG, J. AND GHOSE, K. 2012. Detecting and Tracking Coordinated Groups in Dense, Systematically Moving, Crowds. In *SDM '12*. 1–11.

SIMPLY HIRED, INC. Portland Jobs. Last accessed May 16 2012 http://www.simplyhired.com/a/local-jobs/city/l-Portland,+OR.

TAKAHASHI, R., OSOGAMI, T., AND MORIMURA, T. 2012. Large-Scale Nonparametric Estimation of Vehicle Travel Time Distributions. In *SDM '12*. 12–23.

THE ENGINEERING TOOLBOX. Common Area per Person in Buildings. Last accessed: May 16 2012 http://www.engineeringtoolbox.com/number-persons-buildings-d_118.html.

THE OFFICIAL U.S. GOVERNMENT SOURCE FOR FUEL ECONOMY INFORMATION. Fuel Economy. Last accessed May 16 2012 http://www.fueleconomy.gov/feg/,.

TISHBY, N., PEREIRA, F. C., AND BIALEK, W. 1999. The Information Bottleneck Method. In *37th Annual Allerton Conference on Communication, Control and Computing*. 368–377.

U.S. GENERAL SERVICES ADMINISTRATION. 1997. Office Space Use Review, Current Practices and Emerging Trends.

YANG, J. AND LESKOVEC, J. 2011. Patterns of Temporal Variation in Online Media. In *ACM international conference on Web search and data mining* . 177–186.

YUAN, J., ZHENG, Y., AND ET AL. 2011. Driving with Knowledge from the Physical World. In *KDD '11*.

YUAN, J., ZHENG, Y., AND XIE, X. 2012. Discovering Region of Different Functions in a City Using Human Mobility and POI. In *KDD '12*.

YUAN, J., ZHENG, Y., ZHANG, C., XIE, W., XIE, X., AND HUANG, Y. 2010. T-Drive: Driving Directions Based on Taxi Trajectories. In *ACM SIGSPATIAL GIS 2010*.

ZHENG, Y., ZHANG, L., XIE, X., AND MA, W. 2009. Mining Correlation between Locations Using Human Location History. In *ACM SIGSPATIAL GIS 2009*.

Received October 2012; revised June 2013; accepted August 2013