ICFinder: A Ubiquitous Approach to Detecting Illegal Hazardous Chemical Facilities with Truck Trajectories

Zheng Zhu[1,2], Huimin Ren[1,3], Sijie Ruan[1,4], Boyang Han[1], Jie Bao[1*],
Ruiyuan Li[5,1], Yanhua Li[3], Yu Zheng[1,4*]
[1]JD Intelligent Cities Research [2]University of Electronic Science and Technology of China, Sichuan, China
[3]Worcester Polytechnic Institute, Worcester, MA, USA [4]Xidian University, Shaanxi, China
[5]College of Computer Science, Chongqing University, China
{zhengzhu97,sijieruan,msyuzheng}@outlook.com;{hren,yli15}@wpi.edu;{hanboyang,baojie}@jd.com;liruiyuan@whu.edu.cn;

ABSTRACT

Chemical materials are useful but sometimes hazardous, which requires strict regulation from the government. However, due to the potential economic benefits, many illegal hazardous chemical facilities are running underground, which poses a significant public safety threat. However, the traditional solutions, e.g., on-field screening and the anonymous tip-offs, involve a lot of human efforts. In this paper, we propose a ubiquitous approach called ICFinder to detecting illegal chemical facilities with chemical transportation trajectories. We first generate candidate locations by clustering stay points extracted from trajectories, and filter out known locations. Then, we rank those locations in suspicion order by modeling whether it has the loading/unloading events. ICFinder is evaluated over the real-world dataset from Nantong in China, and the deployed system identified 20 illegal chemical facilities in 3 months.

CCS CONCEPTS

• **Information systems → Spatial-temporal systems**.

KEYWORDS

hazardous chemicals transportation, trajectory data mining

ACM Reference Format:
Zheng Zhu, Huimin Ren, Sijie Ruan, Boyang Han, Jie Bao, Ruiyuan Li, Yanhua Li, and Yu Zheng. 2021. ICFinder: A Ubiquitous Approach to Detecting Illegal Hazardous Chemical Facilities with Truck Trajectories. In *Proceedings of Beijing '21: 29th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (SIGSPATIAL '21).* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3474717.3483633

*Jie Bao and Yu Zheng are corresponding authors.
The work was supported by the National Key R&D Program of China (2019YFB2101805), NSFC (61976168, 62076191), NSF grants IIS-1942680 (CAREER), CNS-1952085, CMMI1831140, and DGE-2021871.

SIGSPATIAL '21, November 2-5, 2021, Beijing, China

ACM ISBN 978-1-4503-8664-7/21/1…$15.00
https://doi.org/10.1145/3474717.3483633

1 INTRODUCTION

Chemical materials are widely used in our daily lives, such as energy consumption and manufacture industries. Some of the chemicals are very dangerous, e.g., flammable or corrosive. As a result, the processing, transportation, and storage of these hazardous chemical materials are strictly regulated by the government[5]. However, motivated by enormous economic profits, there are many underground and illegal chemical businesses going on, where their storage and processing facilities are not registered and lack of proper protection measures. The mismanagement of the chemicals in these facilities pose a great danger to public safety.

Existing methods to find these illegal chemical facilities either rely on the anonymous tip-off reports, or are based on the active on-field screening over some suspicious areas. However, both of them are time-consuming and achieve limited coverage.

Fortunately, hazardous chemical materials are highly regulated, and allowed to be transported using only the hazardous chemical transportation (HCT) trucks. These trucks are equipped with GPS modules, and the GPS data is reported to the governments directly. Inspired by the fact that HCT trucks trajectories reflect the delivery activities at the chemical facilities, in this paper, we propose ICFinder, a ubiquitous and cost-effective method to detect the illegal hazardous chemical facilities based on HCT trajectories.

ICFinder contains two parts: 1) *Candidate Location Discovery*, which discovers candidate locations based on stay points extracted from HCT trajectories; 2) *Illegal Facility Detection*, which ranks candidate locations based on the probabilities of loading/unloading events. Our main contributions are as follows:

- We propose a novel and ubiquitous way to locate the illegal hazardous chemical facilities using HCT trajectories.
- We propose a two-step approach, which first generates candidate locations based on stay points in trajectories, then models whether each of them has the loading/unloading events based on the spatio-temporal and other contextual features.
- The proposed method is evaluated extensively over the real-world HCT trajectories. The system is deployed in Nantong, and has identified 20 illegal chemical facilities in three months.

2 OVERVIEW

2.1 Definitions and Problem Statement

Definition 1. Trajectory. A trajectory is a sequence of spatio-temporal points, denoted as $tr = \langle p_1, p_2, ..., p_n \rangle$, where each point $p = \langle lat, lng, t \rangle$ consists of a location at time t.

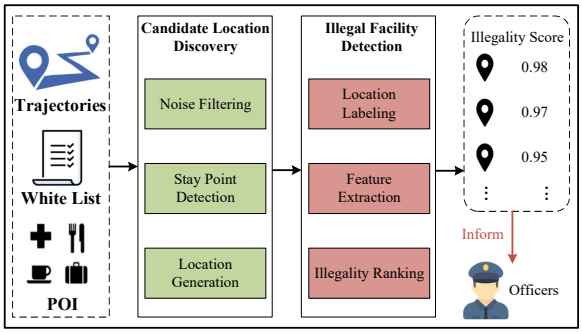


Figure 1: System Framework of ICFinder.

Definition 2. Stay Point. A stay point is a sub-sequence of a trajectory, indicating that the moving object stays in the geographic region for a period of time [4].

Definition 3. Candidate Location. A candidate location is a spatial point, which is the centroid of a group spatially correlated stay points. The location is defined as $l = \langle lat, lng \rangle$.

Definition 4. White List. The white list is a list of facilities, where hazardous chemicals can be legally manufactured and processed (denoted as $\mathcal{W}$). The type of a facility in $\mathcal{W}$ can be classified into two categories: 1) consumer; and 2) producer.

Consumers can legally consume the hazardous chemicals, and are denoted as C. Producers can legally produce hazardous chemicals, and are denoted as $\mathcal{P}$. $\mathcal{W} = C \cup \mathcal{P}$. Each facility in W contains the following properties: 1) name of the facility; 2)permitted hazardous chemical list (producer only); and 3) location.

Definition 5. Illegal Hazardous Chemical Facility. Illegal hazardous chemical facilities consume hazardous chemicals without the proper registration.

Problem Statement: Illegal Hazardous Chemical Facility Detection. Given trajectories of chemical transportation $\mathcal{T}$ and a white list $\mathcal{W}$, we aim to find the locations of illegal chemical facilities.

The system framework of ICFinder is illustrated in Figure 1, which consists of two main components: *candidate location discovery* and *illegal facility detection.*

3 CANDIDATE LOCATION DISCOVERY

In this component, we extract the candidate locations from stay points of HCT trajectories which are stored in JUST [3]. The main intuition here is that HCT trucks must stop, when loading/unloading hazardous chemicals. Thus, the stay points are highly related to illegal chemical facility locations.

In this section, we first process the trajectories with **noise filtering** techniques to remove the outlier points due to the shifts introduced in the GPS sensing. In our model, the point will be filtered out, if its speed is higher than 130 km/h. After that, we **extract the stay points** from the trajectories using a spatial clustering algorithm. In this step, we identify all the stay points from the HCT trajectories by following algorithm [4]. We tried different parameter combinations and found that the most of stay points scattered within the range of 100 meters and 15 minutes. Finally, the stay points are **clustered to generate the candidate locations**. In this way, we can filter out all the random stops during the transportation. A candidate location is discovered from stay points via DBSCAN, since staying area might be in different scale. The centroid of a cluster is the position of the candidate location. In this work, we set $\epsilon = 50m$, $minPts = 8$ in DBSCAN.

4 ILLEGAL FACILITY DETECTION

In this component, we apply a machine learning model to detect if a candidate location is an illegal hazardous chemical facilities.

The most straightforward solution to tackle the problem is classifying the candidate locations directly. However, only 22 illegal chemical facilities were detected in history, which makes it impossible to train the model with such limited labels. On the other side, there usually are some loading/unloading events at the chemical facilities. Therefore, to detect the illegal chemical facilities, we only need to find all the candidate locations, which have loading/unloading events and are not on the white list.

To this end, we first label the candidate locations, based on the white list and the domain knowledge. Then, we train a model to predict if an uncertain location has the L/U events.

4.1 Location Labeling

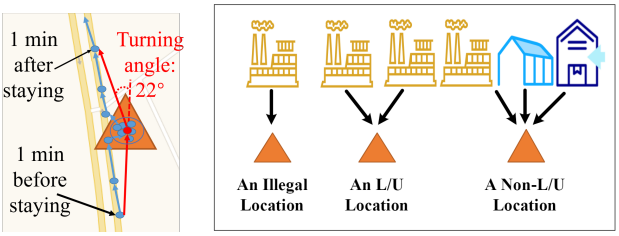
In this step, we first categorize the locations into three groups: 1) white-listed L/U Locations; 2) Non-L/U Locations; and 3) Uncertain Locations. The details of each group are as follows:

- *White-listed L/U location*, which is a candidate location that has the loading/unloading events near a legal chemical facility. A candidate location is labeled as the *white-listed L/U location* if it is within 50 meters of a location in the white list.
- *Non-L/U location*, which is a candidate location that does not have loading/unloading events. Based on the domain knowledge, it is not possible to have any chemical loading/unloading events near some acceptable POI types, e.g., parking lots, toll stations, and service areas. As a result, a candidate location is labeled as the *non-L/U location* if there is an acceptable POI within 50 meters.
- *Uncertain location*, which are all the candidate locations that are not labeled by the above heuristics. All the potential illegal chemical facilities are included in this set.

The first two groups of locations are used as positive and negative labels to train the L/U locations detection model. All the uncertain locations are used for illegal chemical facility detection.

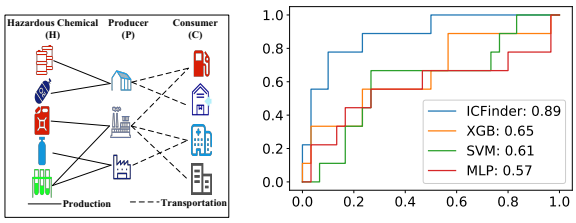
4.2 Feature Extraction

In this step, we extract representative features to identify candidate locations with loading/unloading events. However, there is a challenging, as both legal and illegal chemical locations have L/U events, but many of their features are very different. Therefore, it will be inaccurate, if we use the features, which are only valid at the legal chemical locations (or white-listed L/U locations), to detect the loading/unloading events at the uncertain locations (i.e., illegal chemical facilities). For example, illegal chemical locations are more likely to locate in rural areas compared with legal chemical locations, and illegal chemical locations tend to be visited by fewer HCT trajectories according to their scale.



(a) Turning Angle. (b) Origination Comparison.

Figure 2: Intuitions for Feature Extraction.



(a) Hazardous Transport Graph. (b) Results.

Figure 3: Figures.

To this end, we should only select the features that are shared at both the legal and illegal chemical facilities, but, at the same time, are very different from the Non-L/U locations. Based on our observation, the key difference here is that, comparing to the Non-L/U locations, legal and illegal chemical facilities are destinations of a transportation and consume meaningful chemical combinations. Therefore, we select the features as follows:

Truck Behavior Features. To identify a destination in a trip, we extract the stay duration and turning angles, as follows:

- **Stay Duration.** It is quite intuitive a L/U event usually requires a significant amount of time to proceed. Thus, we extract the average, standard deviation, 20% and 50% percentiles value of stay duration to characterize the stay temporal behavior of HCT trucks at the candidate location.
- **Turning Angles.** It reflects the intention of the stay if it is a destination. Intuitively, HCT trucks will take a U-turn and go back after loading/unloading chemicals. A turning angle is calculated by connecting 1)the GPS points that are one minute before the stay point; 2) the centroid of stay point; and 3) the GPS point that is one minute after the stay point, as shown in Figure 2(a). We extract the feature by averaging the turning angles of all stay points at the candidate location to indicate if the visits are intentional.

Context Features. Based on the domain knowledge, all of the chemical facilities, whether legal or not, require limited combinations of hazardous chemicals. Thus, only limited types of hazardous chemicals will be delivered to L/U locations, while there will be multiple types of hazardous chemicals showing in non-L/U locations. However, we do not know what kind of chemicals are delivered in each HCT trajectory and we do not know whether the HCT trucks contain chemicals or not when stopping by the non-L/U locations. Only the types of permitted chemicals are provided for each producer in L/U locations. Therefore, we need to infer the hazardous chemicals distribution for L/U locations and non-L/U locations via HCT trajectories .

(1) **Latent Chemicals Distribution in L/U Location**: In this step, we learn the latent related chemicals for the producers and consumers. First, we leverage the permitted chemicals from producers, and the transportation relation between producers and consumers to construct a heterogeneous graph, i.e., Hazard Chemicals Transportation Graph (HCTG), as shown in Figure 3(a). There are three types of nodes in HCTG: hazardous chemicals $\mathcal{H}$, producers $\mathcal{P}$, and consumers $\mathcal{C}$. $\mathcal{P}$ and $\mathcal{C}$ are all derived from the white list, and $\mathcal{H}$ is the set of all hazardous chemicals that have appeared in all the producers. The edges between hazardous chemicals and producers are constructed if a producer can produce a specific chemical, and the weight is set to 1. The edges between producers and consumers are constructed by HCT trajectories. If the HCT trucks stay at a producer and a consumer consecutively, we add a heterogeneous edge between the producer and the consumer. The weight of the edge is set according to the total number of trips from the producer to the consumer. Then, we apply a heterogeneous graph representation learning algorithm Metapath2vec [2] to learn the embeddings of each producer and consumer in the graph, which reflects its latent related chemicals.

(2) **Latent Chemicals Distribution in non-L/U Location**: The HCT trucks carry chemicals from facilities in the white list to a non L/U location to have a rest. Thus, we aggregate by weighted average the embeddings of L/U facilities that each truck has stayed before coming to the current location to capture the latent chemical distribution in the non-L/U locations

4.3 Candidate Location Ranking

After extracting features of candidate locations, we train a binary classification model to detect the L/U locations and then rank the probabilities of uncertain locations to detect illegal facilities. Since the number of candidate locations is limited, Xgboost [1] is employed as the classification model instead of deep learning approaches. During the training phase, we employ L/U locations as the positive labels and Non-L/U locations as the negative labels. During the actual deployment, we predict the probability of L/U locations for all uncertain locations and report the top candidate locations to the government.

5 EXPERIMENT EVALUATION

In this section, we first describe the real dataset used in the experiments. After that, we design two experiments to evaluate the effectiveness of the ICFinder: 1) comparing the ICFinder with other traditional end-to-end methods on the same dataset; 2) an on-field case study to show the effectiveness of our ICFinder.

5.1 Data Descriptions

The datasets used in the experiments are all collected from the city of Nantong, China. The detailed descriptions are as follows:

HCT Trajectories. HCT trajectory datasets contain 59.74 million GPS points generated by 2,891 HCT trucks in the City of Nantong from June 1^{th} to October 31^{th}, 2020. Each trajectory contains a

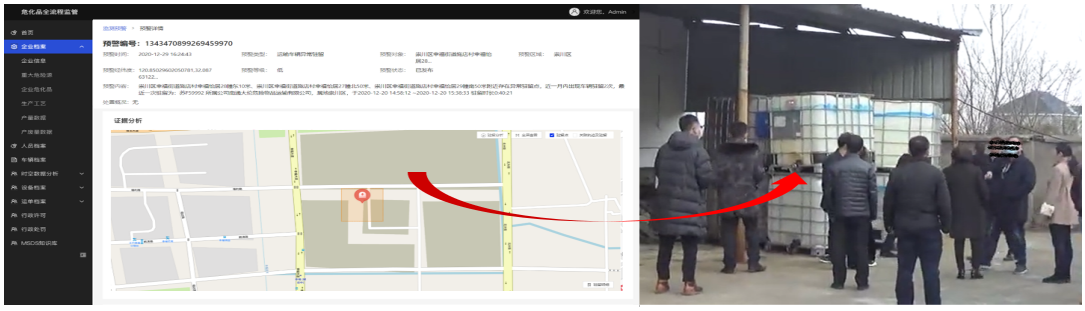


Figure 4: Case Study.

truck ID, a series of GPS points and the corresponding timestamp. The average sampling time interval is around 2.5 minutes.

White List. The white list contains 8,512 facilities with 2,011 producer facilities and 6,501 consumer facilities. The total number of hazardous chemical items is 875.

POI. The POI dataset contains a total number of 415,639 POIs and is categorized into 22 different types.

Ground Truth Labels. 22 ground truth labels are provided by the government, which were found by on-field screening operations.

5.2 Evaluation of Illegal Facility Detection

In order to evaluate the effectiveness of illegal facility detection, we compare our ICFinder with straightforward solutions which are directly trained with the limited chemical facilities.

Dataset. There are 22 illegal chemical facilities. 11 of them are considered as positive labels in the training dataset and others of them are used as the test dataset to evaluate models. Due to the limited number of positive labels, 50 Non-L/U locations and 20 Non-L/U locations are randomly selected as negative labels for the training and test dataset respectively. L/U locations are not selected as negative labels to train baseline models since locations from the white list will be removed during the test. Thus, to ensure the training and test data are in the same distribution, we only use Non-L/U locations as negative labels.

Evaluation Metric. For this imbalance binary classification task we evaluate the performance by AUC, which reflects model performance with different discrimination thresholds.

Baselines. To evaluate the performance of our model, we compare ICFinder with MLP, SVM, and XGB. Different from ICFinder which is trained with truck behavior and context features, we provide baseline models with more spatial-temporal features to capture the location information and truck behavior. The following features are extracted:

- Spatial Features: 1) POI category distribution; 2) total length of road network; and 3) number of road intersections.
- Temporal Features: 1) the average of stay points per day; 2) the average number of unique drivers per day; 3) the mean and variance of vehicles staying time, and 4) the average turning angles of vehicles.

Results. Figure 3(b) illustrates the ROC curves and AUC of our ICFinder compared with other baseline models. Even if the ICFinder is not directly trained on illegal chemical facilities, our method is 0.24 higher in AUC than the best baseline model with XGB. Baseline models perform worse because the number of illegal chemical facilities is limited and complicated features cannot be learned well with such limited data. The results indicate that we cannot solve this problem in a straightforward feature engineering method and loading/unloading pattern detection is an effective way to detect illegal chemical facilities.

5.3 Case Study

We further give a case study to show the effectiveness of our proposed method. On Dec. 29th 2020, a suspicious location was found in Rugao Area, Nantong via ICFinder. As shown in the left part of Figure 4, we found a candidate location (rank score: 0.981), which was consisted of 27 stay points but not recorded on the white list. After notified of this anomaly information, officers arrived at the certain location and verified that it was an illegal chemical facility. The right part of Figure 4 shows a large number of the hazardous chemicals were illegally stored in the warehouse.

6 CONCLUSION

In this paper, we propose a novel approach named ICFinder to find out unregistered and unqualified hazardous chemical facilities by mining HCT trajectories. To avoid redundant candidate locations, we cluster stay points in HCT trajectories into locations and filter out some known locations according to the white list and POIs. To exclude locations with other stay reasons, a machine learning model is then applied to detect illegal chemical facilities. To overcome the issue of label scarcity, we convert the training target of the model from illegal facilities detection into loading/unloading patterns inference. A real-world application system has been deployed in Nantong, China since Nov. 2020. With the help of ICFinder, extra 20 illegal chemical facilities have been identified by our system in the same area, which have been validated by local experts.

REFERENCES

[1] Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD*. 785–794.
[2] Yuxiao Dong, Nitesh V Chawla, and Ananthram Swami. 2017. metapath2vec: Scalable representation learning for heterogeneous networks. In *Proceedings of the 23rd ACM SIGKDD*. 135–144.
[3] Ruiyuan Li, Huajun He, Rubin Wang, Yuchuan Huang, Junwen Liu, Sijie Ruan, Tianfu He, Jie Bao, and Yu Zheng. 2020. Just: Jd urban spatio-temporal data engine. In *ICDE*. IEEE, 1558–1569.
[4] Sijie Ruan, Zi Xiong, Cheng Long, Yiheng Chen, Jie Bao, Tianfu He, Ruiyuan Li, Shengnan Wu, Zhongyuan Jiang, and Yu Zheng. 2020. Doing in One Go: Delivery Time Inference Based on Couriers' Trajectories. In *Proceedings of the 26th ACM SIGKDD*. 2813–2821.
[5] Jingyuan Wang, Chao Chen, Junjie Wu, and Zhang Xiong. 2017. No longer sleeping with a bomb: a duet system for protecting urban safety from dangerous goods. In *Proceedings of the 23rd ACM SIGKDD*. 1673–1681.