CrowdWaterSens: An Uncertainty-aware Crowdsensing Approach to Groundwater Contamination Estimation

Lanyu Shang[a], Yang Zhang[a], Quanhui Ye[b], Shannon L. Speir[c], Brett W. Peters[e], Ying Wu[d], Casey J. Stoffel[e], Diogo Bolster[d,e], Jennifer L. Tank[e,f], Danielle M. Wood[e,g], Na Wei[b], Dong Wang[a,*]

[a]*School of Information Sciences, University of Illinois at Urbana-Champaign, Champaign, IL 61820, United States*
[b]*Department of Civil and Environmental Engineering, University of Illinois at Urbana-Champaign, Urbana, IL 61801, United States*
[c]*Department of Crop, Soil, and Environmental Sciences, University of Arkansas, Fayetteville, AR 72701, United States*
[d]*Department of Civil and Environmental Engineering and Earth Sciences, University of Notre Dame, Notre Dame, IN 46556, United States*
[e]*Environmental Change Initiative, University of Notre Dame, Notre Dame, IN 46556, United States*
[f]*Department of Biological Sciences, University of Notre Dame, Notre Dame, IN 46556, United States*
[g]*Center for Civic Innovation, University of Notre Dame, Notre Dame, IN 46556, United States*

Abstract

Groundwater contamination poses serious threats to public health and environmental sustainability. In this paper, we explore *smart groundwater contamination sensing*, which aims to accurately estimate the nitrate concentration in groundwater via a crowdsensing approach. Existing solutions often require professional groundwater collection and high-quality measurement of groundwater properties, making the data collection process time-consuming and unscalable. In this work, we leverage the approximate nitrate concentration measured by crowd sensors (i.e., participants from well-dependent communities) to accurately estimate nitrate concentration in groundwater samples. Three critical challenges exist in developing the crowdsensing-based groundwater contamination estimation solution: i) the spatial irregularity of the crowdsensing groundwater contamination data, ii) the hidden temporal dependency of ground-

*Corresponding author

Email addresses: lshang3@illinois.edu (Lanyu Shang), yzhangnd@illinois.edu (Yang Zhang), quanhui2@illinois.edu (Quanhui Ye), slspeir@uark.edu (Shannon L. Speir), bpeters2@nd.edu (Brett W. Peters), ywu10@nd.edu (Ying Wu), cstoffe2@nd.edu (Casey J. Stoffel), bolster@nd.edu (Diogo Bolster), tank.1@nd.edu (Jennifer L. Tank), dwood5@nd.edu (Danielle M. Wood), nawei2@illinois.edu (Na Wei), dwang24@illinois.edu (Dong Wang)

water contamination in the anthropogenic context, and iii) the uncertainty of crowdsensing nitrate measurements from crowd sensors. To address the above challenges, we develop CrowdWaterSens, an uncertainty-aware graph neural network framework that explicitly examines the uncertainty and spatial irregularity of the crowdsensing groundwater contamination data and its relevant anthropogenic context to accurately estimate groundwater nitrate concentration. We evaluate the CrowdWaterSens framework through two real-world case studies in well-dependent communities in Northern Indiana, United States. The evaluation results not only show the effectiveness of CrowdWaterSens in accurately estimating nitrate concentration, but also demonstrate the viability of crowdsensing for community-level groundwater quality monitoring.
Keywords: Groundwater Quality, Nitrate Contamination, Crowdsensing, Graph Neural Network

1. Introduction

Groundwater is one of the critical natural resources on Earth [1]. For example, more than 115 million people in the United States rely on groundwater as their primary drinking water source.[1] However, groundwater resources are vulnerable to contamination induced by various human activities, such as excessive application of fertilizer and pesticide in agricultural operations, failure of private septic systems, and uncontrolled waste disposal on abandoned dumpsites and hazardous waste sites [1]. Groundwater contamination poses serious threats to public health and the environment [2]. Among groundwater contaminants, nitrate (NO_3^--N) is an important widespread contaminant that can cause serious health issues [3]. For example, long-term intake of groundwater with elevated nitrate concentration can cause methemoglobinemia (i.e., blue baby syndrome) and stomach cancer [3]. In this paper, we developed a *smart groundwater contamination sensing* approach that aims to accurately estimate the nitrate concentration in groundwater via crowdsensing. The estimated nitrate concentration results can be further reported to federal and local groundwater quality monitoring agencies (e.g.,

[1] https://www.usgs.gov/media/images/groundwater

U.S. Environmental Protection Agency, state department of ecology/health) to aid in improving groundwater quality.

Our work is motivated by the lack of effective government regulations on well-water quality monitoring in the United States and the lack of consistent testing for groundwater contamination in well-dependent communities (i.e., households consuming groundwater from private wells) [4]. The quality of groundwater from private wells is not regulated by federal or state laws (e.g., Federal Safe Drinking Water Act) in the United States [5] and it is typically homeowner's responsibility to maintain their private well systems and monitoring the groundwater quality. However, due to general lack of knowledge, residents in well-dependent communities are often unaware of the health risk posed by groundwater contamination and the importance of groundwater quality monitoring [4]. Therefore, it is necessary to develop effective groundwater quality monitoring solutions to accurately estimate the contamination in private well water while increasing the awareness of groundwater contamination in well-dependent communities.

Recent efforts have been made towards groundwater contamination estimation [6]. Existing solutions mainly focus on the spatial interpolation of groundwater contamination based on geographic information system (GIS) data [7] and groundwater properties (e.g., hydrochemical facies [8], geochemical and microbiological features [9]) measured at groundwater sampling sites (e.g., groundwater monitoring stations, designated private wells). However, such methods often require professional groundwater collection from well-established sampling sites and high-quality measurements of groundwater properties, making the data collection process costly, time-consuming, and unscalable. Therefore, it remains challenging to effectively and efficiently monitor groundwater contamination at the scales of relevant interest.

In this paper, we develop a crowdsensing-based approach that explores the collective wisdom of crowd sensors (i.e., participants from well-dependent communities) to accurately estimate groundwater nitrate concentration. Crowdsensing, when implemented correctly, presents an effective data collection paradigm for obtaining measurements from non-technical individuals in an efficient and scalable way [10, 11]. In particular, our crowdsensing-based contamination estimation approach is in principle

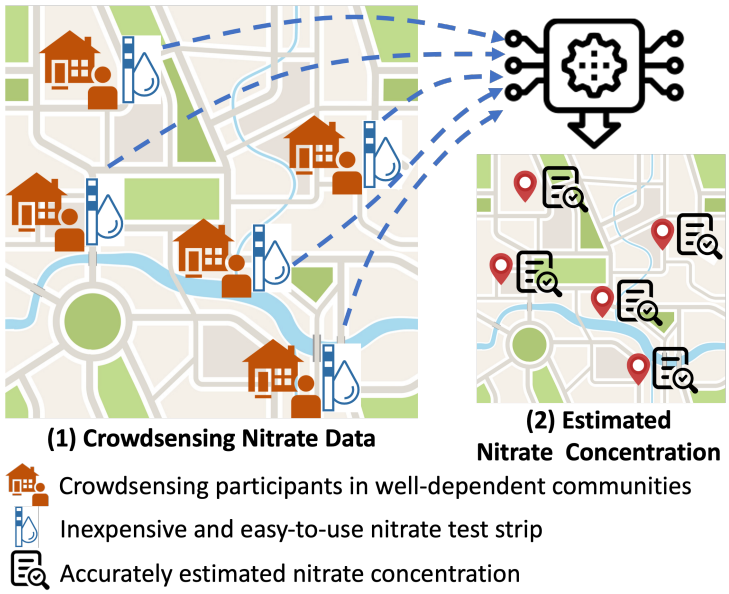


Figure 1: Crowdsensing Groundwater Nitrate Estimation

infrastructure-free as it does not require any installation or maintenance of physical water quality monitoring sensors. Moreover, crowdsensing data collection strategies also actively engage members in well-dependent communities to monitor groundwater quality and thus become informed about the safety of their drinking water. Figure 1 depicts a schematic representation of our crowdsensing-based groundwater nitrate estimation approach. First, we collect the crowdsensing nitrate data (i.e., approximate nitrate concentration using test strips) from crowd sensors. The collected crowdsensing nitrate data is then leveraged to accurately estimate the nitrate concentration in the studied communities. However, several challenges exist in developing our solution.

Spatial Irregularity. The first challenge lies with the spatial irregularity of the crowdsensing groundwater nitrate data. Existing methods for spatial estimation of groundwater contamination often rely on the groundwater properties (e.g., hydrochemical facies) measured from a set of fixed water sampling stations [8]. In addition, these fixed sampling stations are pre-identified based on geographic characteristics (e.g., watershed locations, soil conditions). However, such methods cannot be readily applied to model the crowdsensing groundwater nitrate data which are irregularly located across

study areas and dynamically change due to the random nature of crowdsensing participants. For example, the crowdsensing data from areas with higher residential density are denser than crowdsensing data from areas with lower residential density. This poses a challenge in effectively extracting the spatial relations among the irregular crowdsensing groundwater nitrate data.

Hidden Temporal Dependency. The second challenge lies in the hidden temporal dependency of groundwater contamination on the anthropogenic context (i.e., human activities and land use that can cause groundwater contamination). The concentration of groundwater nitrate is often correlated with the temporal dynamics of anthropogenic activities (e.g., temporal variation of fertilizer application and pet activities), especially for groundwater obtained from shallow wells in well-dependent communities [4]. For example, the application of fertilizer is often dependent on individual households (e.g., professionally maintained landscape vs. poorly maintained landscape) and the season of the year (e.g., summer vs. winter). The varying dynamics of fertilizer activities can affect the nitrate concentration in groundwater [12]. Such variability poses a challenge to efficiently capture hidden temporal dependencies to accurately estimate groundwater nitrate concentration.

Crowdsensing Data Uncertainty. The third challenge lies in the uncertainty of the crowdsensing nitrate data collected from the crowd sensors who are often untrained citizens from well-dependent communities. Crowdsensing nitrate concentrations measured by such non-professional crowd sensors can present significant uncertainties due to the unknown reliability of the participants, arising from the lack of familiarity with the nitrate test kit and incorrect sampling of groundwater. While several data uncertainty estimation solutions exist (e.g., truth discovery [13, 14], data reliability assessment [15, 16]), such approaches cannot be directly applied to address data uncertainty issues in our crowdsensing nitrate estimation problem. This is because these approaches typically assume that multiple observations/measurements of the target variable (e.g., nitrate concentration) from different sources (e.g., participants) are available or they require a large number of data samples with annotated reliability (e.g., reliability of the participants). However, in our problem, each participant is only asked to measure the nitrate concentration from his/her own household, and only a limited

number of ground truth nitrate concentration results are collected to assess the participants' reliability due to the high cost of professional nitrate measurements. Therefore, it remains a challenge to accurately predict groundwater nitrate concentration based on crowdsensing measurements from participants with unknown reliability.

In this paper, we develop CrowdWaterSens, an uncertainty-aware crowdsensing framework to accurately and efficiently estimate groundwater nitrate concentration by exploring the collective accuracy of well-dependent communities. To address the first challenge, we design a spatio-temporal crowdsensing network to explicitly model the spatial irregularity of the crowdsensing groundwater nitrate data. To address the second challenge, we develop a context-aware information fusion module to effectively learn the temporal dynamics of crowdsensing contamination and its relevant anthropogenic context to jointly estimate the groundwater nitrate concentration. To address the third challenge, we design an uncertainty-driven network optimization strategy that explicitly measures the uncertainty of crowdsensing measurements from individual crowd sensors to accurately estimate the groundwater nitrate concentration via a graph-based regression model. To the best of our knowledge, CrowdWaterSens is the first uncertainty-aware crowdsensing-based groundwater contamination estimation solution for groundwater quality monitoring. We evaluate the proposed CrowdWaterSens framework by carrying out two case studies of nitrate contamination in well-dependent communities in Northern Indiana, United States.[2] The evaluation results not only show the effectiveness of CrowdWaterSens in accurately estimating nitrate concentration but also demonstrate the viability of crowdsensing for community-level groundwater quality monitoring.

A preliminary version of this work (i.e., SmartWaterSens [17]) has been published in SmartComp 2022 to study the problem of groundwater monitoring in well-dependent communities. This paper is a significant extension of our conference paper (i.e., SmartWaterSens) in the following aspects. First, we identify a new critical challenge of data uncertainty in crowdsensing-based groundwater contamination estimation (Section 1).

[2]The Institutional Review Board (IRB) approval has been granted for all protocols and procedures in our case study.

Second, we extend the preliminary solution by explicitly modeling the crowd sensor reliability in the graph regression framework to address the data uncertainty challenge in crowdsensing-based groundwater nitrate concentration estimation. Third, in light of the COVID-19 pandemic, we extend our crowdsensing experiments to a new case study with a fully at-home crowdsensing data collection strategy that obtains the crowdsensing measurements and ground truths of nitrate concentration, while minimizing physical contact with our participants (Section 5). Fourth, we add three new baselines of graph neural network based regression models, GCN, GTN, and SmartWaterSens to further investigate the effectiveness and efficiency of the proposed CrowdWaterSens framework. Fifth, we carry out a robustness study to comprehensively evaluate the effect of training data size on the performance of CrowdWaterSens. Last, we extend the related work discussion by adding a discussion on the topic of data uncertainty in crowdsensing and including more recent work in groundwater quality, crowdsensing, and spatio-temporal inference (Section 2).

2. Related Work

2.1. Groundwater Quality

Groundwater quality has gained much attention in recent years [1]. Many efforts have been made to investigate groundwater pollution and assess groundwater quality [6, 18]. For example, Long *et al.* introduced a spatially interpolated mapping system to estimate the spatial distribution and health risk of heavy metals in shallow groundwater [19]. Li *et al.* designed an entropy-weighted multi-criteria decision analysis approach to assess the plateau groundwater quality based on hydrochemical facies (e.g., concentrations of major ions) [20]. Egbueri *et al.* proposed a hierarchical cluster analysis to jointly investigate the pollution index of groundwater (PIG) and ecological risk index (ERI) for drinking groundwater quality assessment [21]. Knoll *et al.* developed a GIS-based machine learning framework that leverages nitrate measurements at groundwater monitoring sites to predict groundwater nitrate concentration [22]. However, existing methods often require well-established infrastructures or laboratory analysis which are time-consuming and expensive. In this paper, we present an infrastructure-free solution that explores the crowdsensing wisdom from citizen scientists to assess

the quality of groundwater.

2.2. Crowdsensing in Smart City Applications

Crowdsensing presents a new sensing paradigm, where timely observations of the physical world are collected from human sensors [23, 24]. With the pervasive network connections and the prevalence of digital devices, crowdsensing has been increasingly applied in smart city applications [25, 26, 27]. For example, Liang *et al.* leveraged crowdsensing data from public air quality sensors to assess the wildfire smoke impact of indoor air quality in California [28]. Silva *et al.* designed a crowd-driven vehicle pollution monitoring system that couples crowdsensing with an on-board diagnostic carbon dioxide reader to estimate vehicle emission in smart cities [29]. Zhang *et al.* proposed a multi-view learning framework to identify risky traffic locations in smart transportation systems [30]. Breuer *et al.* developed HydroCrowd, a crowdsourcing-based water sampling strategy that only recruited crowd participants to collect surface water samples for hydrological study [31]. Lee *et al.* proposed a crowdsensing noise mapping framework to monitor urban environmental noise in smart cities by utilizing crowdsourced noise data from calibrated smartphones [32]. To the best of our knowledge, CrowdWaterSens is the first uncertainty-aware crowdsensing approach to estimate groundwater contamination by leveraging crowdsensing contamination measurements.

2.3. Spatio-Temporal Inference

Our work is also related to spatio-temporal inference that jointly exploits spatial and temporal information in a variety of applications [33]. Examples of spatio-temporal inference applications includes urban traffic monitoring [34], location-based activity prediction [35], climate and weather forecasting [36]. For example, Luo *et al.* developed a deep learning solution to predict urban traffic flows by leveraging spatial traffic flow information extracted from k-nearest neighbor (KNN) monitoring stations and the temporal traffic flow variability learned with a long short-term memory network [37]. Liu *et al.* designed a generative neural network framework for personalized point-of-interest (POI) recommendation using the temporal location-based social network data [38]. Castro *et al.* proposed a spatio-temporal convolutional sequence to

sequence network that models the historical weather records as temporal sequences of spatial grids for temperature and rainfall prediction [39]. However, existing spatio-temporal inference solutions often rely on a large amount of historical and spatial data for accurate estimation. Therefore, these solutions are insufficient to address our crowdsensing-based groundwater nitrate estimation problem where the historical and spatial nitrate contamination data are not always available, especially in rural areas. By contrast, CrowdWaterSens designs a graph-based context-aware spatial-temporal inference model that is dedicated to estimating groundwater nitrate concentration with sparse crowdsensing data.

2.4. Data Uncertainty in Crowdsensing

Data uncertainty is a fundamental problem in the data analysis of crowdsensing systems, where the data are collected from the observation or measurements of ordinary human sensors who are not as reliable as the professionally trained experts [40, 41]. A number of solutions have been proposed to address the data uncertainty issue in crowdsensing systems [42, 43, 16, 14]. For example, Lan *et al.* designed a machine learning based crowdsourcing quality prediction framework that leverages an advanced ensemble machine learning classification algorithm to detect cheating workers on crowdsourcing tasks [16]. Probert *et al.* developed a linguistic model to estimate the uncertainty of crowdsourced data of alien species of birds [44]. Such solutions often require a non-trivial amount of ground truth labels of the crowdsensing measurements to supervise the training of the data uncertainty estimation model. However, they cannot be applied to our groundwater monitoring problem, where only a relatively small number of ground truth nitrate measurements are available due to the expensive cost of professional measurements of nitrate concentration in groundwater samples. In addition, several unsupervised crowdsensing data uncertainty estimation solutions also exist. Wang *et al.* developed a maximum likelihood estimation algorithm approach to estimate crowd sensor reliability in crowdsensing applications [45]. Zhang *et al.* proposed a quality-aware disaster damage assessment framework that integrates estimation theory with deep learning to obtain an accurate assessment of disaster damage severity using crowdsourcing inputs for effective disaster response in natural disaster

events [13]. Such frameworks often assume the existence of multiple observations of the same measured variable (e.g., nitrate concentration) from different sources (e.g., crowd sensors). However, they are insufficient to address data uncertainty in our problem, where each crowd sensor is solely responsible for measuring the groundwater concentration in his/her own household. To address the crowdsensing data uncertainty challenge, we design an uncertainty-driven network optimization strategy to explicitly model the uncertainty of crowdsensing measurements from individual crowd sensors to optimize the estimation performance of groundwater nitrate concentration.

3. Problem Definition

In this paper, we focus on the problem of estimating groundwater nitrate concentration using crowdsensing nitrate measurements from community participants. We first define a few key concepts that will be used in our problem definition.

Definition 1. Participant (p): We define a participant as a person who lives in a well-dependent community and is engaged to report crowdsensing measurements of nitrate (see Definition 5 below). In particular, we define a set of K participants denoted by $P = \{p_1, p_2, \cdots, p_K\}$. We assign each participant a unique anonymized identifier and strictly follow data security precautions to protect participants' privacy.

Definition 2. Participating Community (u): We define a participating community u as the well-dependent community where the households rely on groundwater from private wells as their primary water resource. We define a set of J participating communities as $U = \{u_1, u_2, \cdots, u_J\}$. All the participants in our study are recruited from U.

Definition 3. Location (l): The location $l_k = (\varphi_k, \lambda_k)$ refers to the geographic coordinates at the household of participant p_k, where φ_k and λ_k are the latitude and longitude coordinates, respectively. The collected geolocation information is used for research purpose only and will not be disclosed to unauthorized individuals and third parties.

Definition 4. Sensing Cycle (t): The sensing cycle is the collection period (e.g., daily, weekly) during which the crowdsensing measurements are measured by the participants. In particular, we define T as the total number of sensing cycles in our study and t is the t^{th} sensing cycle.

Definition 5. Crowdsensing Measurement (c): We define the crowdsensing measurement as the test strip reading of the nitrate concentration from a tap water sample (measured in mg/L) using the nitrate test kit. More details about the nitrate test kit will be discussed in Section 5. For each participant $p_k \in P$, we define a set of T crowdsensing measurements reported by p_k as $C_k = [c_{k,1}, c_{k,2}, \cdots, c_{k,T}]$, where $c_{k,t} \in C_k$ represents the crowdsensing measurement measured at location l_k and sensing cycle t. The set of crowdsensing measurements from K participants is denoted as $C = \{C_1, C_2, \cdots, C_K\}$.

Definition 6. Anthropogenic Context (h): We define the anthropogenic context of a participant's household as a set of human-induced features that are often related to groundwater nitrate concentration [4]. In particular, we focus on the following anthropogenic features.

- **Community Type (h^c):** Community type $h^c \in \{$"urban", "suburban", "rural"$\}$ represents the type of community which the participant's household belongs to.
- **Farm Proximity (h^f):** Farm proximity $h^f \in \{$"Yes", "No", "Unsure"$\}$ indicates whether there is a farm nearby the participant's household.
- **Fertilizer Application (h^e):** Fertilizer application $h^e \in \{$"Yes", "No", "Unsure"$\}$ describes whether fertilizer is applied within one week before the crowdsensing measurement is reported.
- **Pet Activity (h^a):** Pet activity $h^a \in \{$"Yes", "No"$\}$ represents whether any outside pet activity is allowed at the participant's household.

Formally, we define the anthropogenic context of T crowdsensing measurements (i.e., C_k) reported by participant p_k as $H_k = [h_{k,1}, h_{k,2}, \cdots, h_{k,T}]$, where $h_{k,t} = [h^c_{k,t}, h^f_{k,t},$

$h^e_{k,t}, h^a_{k,t}]$ is the anthropogenic context of $c_{k,t} \in C_k$. The set of anthropogenic context at the households of K participants is denoted as $H = \{H_1, H_2, \cdots, H_K\}$.

Definition 7. Ground-truth Concentration (y): We define the ground-truth concentration as the nitrate concentration of a tap water sample measured in the laboratory. For each crowdsensing measurement $c_{k,t} \in C_k$ reported by participant p_k, we denote $y_{k,t}$ as the corresponding ground-truth concentration.

Definition 8. Estimated Concentration ($\hat{y}$): We define the estimated concentration of nitrate in groundwater to be the output nitrate concentration predicted by a nitrate estimation model. For each crowdsensing measurement $c_{k,t} \in C_k$ and anthropogenic context $l_{k,t}$ reported by participant p_k, we denote $\hat{y}_{k,t}$ as the estimated concentration.

With the above definitions, the goal of our crowdsensing-based nitrate concentration estimation problem is to accurately estimate the nitrate concentration in the participating communities by exploring the underlying spatio-temporal correlations between the crowdsensing measurements and the relevant anthropogenic context. Formally, given a set of N crowdsensing measurements C and their relevant anthropogenic contexts H, our problem is formulated as:

$$\arg\min_{\hat{y}_i} \left(\mathcal{F}(\hat{y}_i, y_i) \mid C, H\right) \ \forall\, 1 \leq i \leq N \tag{1}$$

where $\mathcal{F}$ is the error measurement function (e.g., Mean Squared Error (MSE), Mean Absolute Error (MAE)) that measures the difference between the estimated and ground-truth concentrations.

4. Solution

In this section, we present the CrowdWaterSens framework to address the crowdsensing-based nitrate concentration estimation problem. CrowdWaterSens is a *context-aware graph neural network* framework that carefully captures the underlying correlations between groundwater nitrate concentration and anthropogenic context variables to infer the hidden spatial and temporal dynamics of groundwater nitrate concentration for desirable estimation performance. An overview of the CrowdWaterSens framework

is shown in Figure 2. CrowdWaterSens contains three main modules: i) the *Spatio-temporal Crowdsensing Network (SCN)* module that constructs a graph-based crowdsensing contamination network to explicitly extract the spatial and temporal relations of the crowdsensing measurements, ii) the *Context-aware Information Fusion (CIF)* module that designs a principled context-aware information propagation mechanism to jointly fuse the spatial and temporal relation of crowdsensing nitrate concentration and the anthropogenic context in the crowdsensing contamination network, and iii) the *Uncertainty-driven Network Optimization (UNO)* module that develops an uncertainty-driven graph regression model that examines the data uncertainty of the crowdsensing measurements to accurately estimate the nitrate concentration in participating communities.

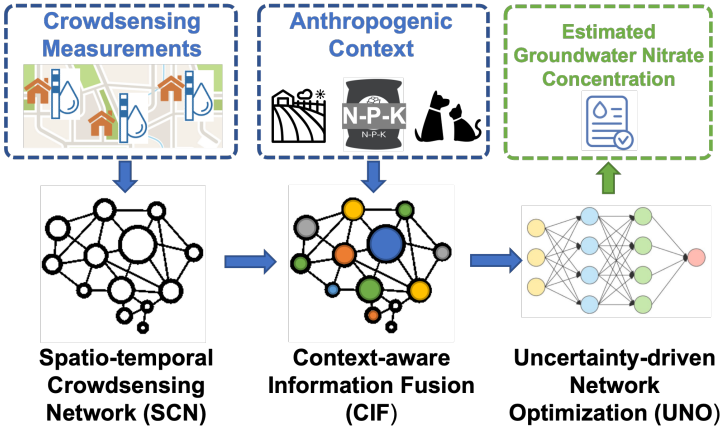


Figure 2: Overview of the CrowdWaterSens Framework

4.1. Spatio-temporal Crowdsensing Network (SCN)

The spatio-temporal crowdsensing network module is designed to construct a graph-based crowdsensing contamination network (CCN) to explicitly model the spatial and temporal relations of the crowdsensing nitrate concentration at different time and locations. Existing nitrate concentration estimation methods mainly rely on geographic grid-based matrices to encode the spatio-temporal information of nitrate concentration. However, such grid-based methods are insufficient to fully capture the spatio-temporal features of crowdsensing data in our problem due to the spatial irregularity of the crowd

participants. To address this limitation, we design a graph-based crowdsensing contamination network (Figure 3) to jointly characterize the spatial and temporal relations of spatial irregular crowdsensing nitrate concentration. Formally, we define the crowdsensing contamination network as below.

Definition 9. Crowdsensing Contamination Network (CCN): An undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is a set of crowdsensing nodes, $\mathcal{E}$ is a set of edges between crowdsensing nodes. In particular, each crowdsensing node $v_{k,t} \in \mathcal{V}$ represents the crowdsensing measurement obtained at location l_k and sensing cycle t. We consider two types of edges in $\mathcal{E}$, including the *spatial edge* $\mathcal{E}_s \in \mathcal{E}$ that represents the spatial distance between two crowdsensing nodes, and *temporal edge* $\mathcal{E}_m \in \mathcal{E}$ that represents the temporal relation between two crowdsensing nodes at the same location.

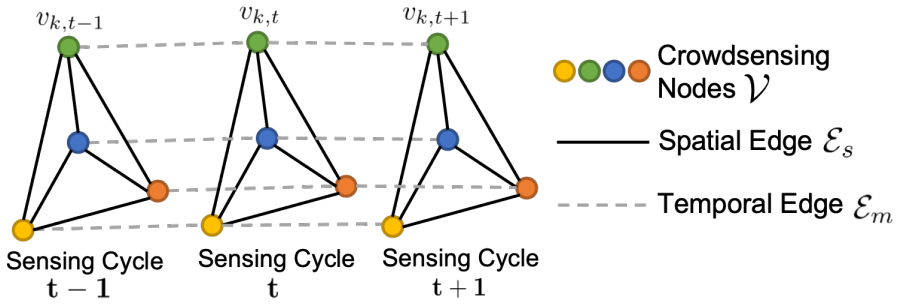


Figure 3: Example of Crowdsensing Contamination Network

We further define the weight of each spatial edge $\mathcal{E}_s$ in CCN to be the physical distance between the locations of two crowdsensing nodes. The assumption is that groundwater samples at nearby locations often have similar nitrate concentrations due to spatial correlation in groundwater settings [12]. In addition, we define the weight of each temporal edge $\mathcal{E}_m$ in CCN as the binary value to indicate the temporal dependency of two crowdsensing nodes (i.e., whether the nitrate concentrations in two crowdsensing nodes are measured in consecutive sensing cycles). The temporal dependency of two crowdsensing nodes that are not measured in consecutive sensing cycles can be captured through the shortest path that includes multiple temporal edges between two nodes. The idea is that groundwater nitrate concentration also changes over time in response to weather events, such as rainfall and drought [46]. To characterize the spatial

and temporal relations between crowdsensing nodes, we define the adjacency matrices to formally represent the spatial and temporal edges and their weights in CCN as below.

Definition 10. Adjacency Matrices ($\mathcal{A}$): $\mathcal{A} = \{S, M\}$ denotes the spatial and temporal adjacent matrices $S \in \mathbb{R}^{N \times N}$ and $M \in \mathbb{R}^{N \times N}$ that correspond to the weights of spatial and temporal edges, respectively. The weight of each spatial edge in S measures the reversed spatial proximity between the locations of two crowdsensing nodes as $1 - \delta(l_1, l_2)$, where $\delta(\cdot)$ is the normalized Haversine distance [47] between location l_1 and l_2. Each value in M is a binary value that indicates whether the nitrate concentrations in two crowdsensing nodes are measured in consecutive sensing cycles (i.e., 1) or not (i.e., 0).

Given the above CCN design, our next step is to learn the temporal dynamics of anthropogenic context of participating communities at different sensing cycles, which is discussed in the next subsection.

4.2. Context-aware Information Fusion (CIF)

The context-aware information fusion module aims to effectively learn the representation of each crowdsensing node in CCN to jointly fuse the spatial and temporal relations of the crowdsensing measurements and anthropogenic contexts (e.g., abnormal rainfall during severe weather and seasonal application of lawn fertilizer) for accurate nitrate concentration estimation. To that end, we collect observations of anthropogenic activities from the crowdsensing participants to explore the spatial and temporal dynamics of the anthropogenic context. In particular, we design a dual-relational graph convolutional network (DR-GCN) to encode the crowdsensing anthropogenic context into CCN to improve the nitrate contamination estimation performance.

DR-GCN is designed to effectively fuse the information of crowdsensing measurements and their anthropogenic context from spatially or temporally related crowdsensing nodes in CCN. Specifically, we first define the encoded vector representation of each crowdsensing node $v_i \in \mathcal{V}$ as:

$$\widetilde{v}_i = [\widetilde{c}_i, \widetilde{h}_i] \in \mathbb{R}^d, \ \forall \ v_i \in \mathcal{V} \tag{2}$$

where $\widetilde{c}_i$ is the normalized crowdsensing measurement (Definition 5) at v_i and $\widetilde{h}_i$ is the one-hot encoding of the anthropogenic context (Definition 6) of v_i. We denote $\widetilde{V} \in \mathbb{R}^{N \times d}$ as the representation matrix of all crowdsensing nodes in $\mathcal{V}$.

Finally, we define the dual-relational aggregation strategy in DR-GCN as:

$$\widetilde{w}_i = \sigma\left(\frac{1}{r_s}\widetilde{S}\widetilde{V}\Theta_s + \frac{1}{r_m}\widetilde{M}\widetilde{V}\Theta_m\right)_i \tag{3}$$

where $\widetilde{w}_i$ is the latent representation of v_i from DR-GCN. $\sigma(\cdot)$ is the non-linear ReLU activation function. $\widetilde{S}$ and $\widetilde{M}$ are the normalized first-order approximation [48] of spatial and temporal adjacent matrices, respectively. r_s and r_m are the normalization constants. Θ_s and Θ_m are learnable parameters.

4.3. Uncertainty-driven Network Optimization (UNO)

Given the context-aware representation of each crowdsensing node learned by DR-GCN, our next goal is to explicitly examine the data uncertainty of the crowdsensing measurements of each crowdsensing node in CCN and accurately estimate the corresponding groundwater nitrate concentration. A key challenge of crowdsensing-based groundwater nitrate estimation is the data uncertainty of crowdsensing nitrate measurements which are measured by ordinary community participants who may not be as reliable as the professionally trained experts due to the lack of sufficient knowledge or experience. For example, a participant may misunderstand the nitrate testing instruction and not wait a sufficient amount of time to read the nitrate concentration results from the test strip, where the test pad color is not stabilized and may lead to an inaccurate nitrate measurement of the groundwater sample. To address such a challenge, we explicitly model the uncertainty of the crowdsensing nitrate measurements contributed by individual crowd sensors. In particular, for each crowd node v_i in CCN, we define the crowdsensing uncertainty ε_i as the variance or the average error of crowdsensing measurements in v_i contributed by the same crowd sensor p_k during the same sensing cycle t. The computing method of crowdsensing uncertainty (i.e., variance or the average error) depends on the design of the crowdsensing experiments which will be discussed in detail in Sections 5 and 6.

While DR-GCN in the CIF module can aggregate the crowdsensing measurements from spatially or temporally correlated crowdsensing nodes in CCN, it ignores the uncertainty of the crowdsensing measurements and considers all crowdsensing measurements as equally important. Such an approach may degrade the nitrate estimation performance when nitrate information from inaccurate crowdsensing measurements is fused in CCN. Therefore, we further learn the uncertainty-aware vector representation of each crowdsensing node in CCN to jointly model the crowdsensing measurement uncertainty and the crowdsensing vector representation learned from the CIF module. In particular, we first update the encoded vector representations $\widetilde{v}_i$ (Equation 2) of each node $v_i \in \mathcal{V}$ to obtain the uncertainty-aware vector representation as:

$$\widetilde{v}_i^* = \frac{1}{\varepsilon_i + \varepsilon_0} \widetilde{v}_i \tag{4}$$

where ε_0 is a normalization factor to avoid the division-by-zero issue. We also denote the uncertainty-aware representation matrix of all crowdsensing nodes in $\mathcal{V}$ as $\widetilde{V}^* \in \mathbb{R}^{N \times d}$. The goal of learning the uncertainty-aware vector representation of each crowdsensing node is to control the crowdsensing measurement error in estimating the nitrate concentration by reducing the effect of crowdsensing measurements from unreliable crowdsensing participants (i.e., the crowdsensing nodes with high crowdsensing uncertainty ϵ_i). Intuitively, the uncertainty-aware vector representation $\widetilde{v}_i^*$ from a crowdsensing node v_i with a high uncertainty ϵ_i is expected to contribute less in estimating the nitrate concentration through the context-aware information fusion and vice versa.

With the uncertainty-aware vector representation, our next objective is to learn the uncertainty-aware latent representations of all crowdsensing nodes in $\mathcal{V}$ with the dual-relational aggregation strategy as presented in Equation 5. In particular, let $\widetilde{V}^* = [\widetilde{v}_1^*, \widetilde{v}_2^*, \cdots, \widetilde{v}_N^*]$ be the matrix of the uncertainty-aware vector representation for all crowdsensing nodes in $\mathcal{V}$. The uncertainty-aware latent representation w_i of node $v_i \in \mathcal{V}$ is defined as below.

$$\widetilde{w}_i^* = \sigma\left(\frac{1}{z_s}\widetilde{S}\widetilde{V}^*\Phi_s + \frac{1}{z_m}\widetilde{M}\widetilde{V}^*\Phi_m\right)_i \tag{5}$$

where $\sigma(\cdot)$ is the non-linear ReLU activation function. $\widetilde{S}$ and $\widetilde{M}$ are the normal-

ized first-order approximation [48] of spatial and temporal adjacent matrices, respectively. z_s and z_m are the normalization constants. Φ_s and Φ_m are learnable parameters. Then, the learned uncertainty-aware latent representations are input into a stacked feed-forward neural network (i.e., multi-layer perceptron) to estimate the corresponding nitrate contamination. Formally, the output of the CrowdWaterSens framework is computed as:

$$\hat{y}_i = \text{MLP}(\widetilde{w}_i) \tag{6}$$

where $\widetilde{w}_i$ is the latent representation of the crowdsensing node $v_i \in \mathcal{V}$, and $\hat{y}_i$ is the estimated concentration. Let y_i be the ground-truth concentration, our learning objective is to minimize the mean squared error loss as:

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2 \tag{7}$$

We adopt the Adaptive Moment Estimation optimizer to learn the accurate value of nitrate concentration in CrowdWaterSens.

5. Crowdsensing Data

In this section, we introduce two case studies we have conducted to collect the crowdsensing groundwater contamination data for the evaluation of CrowdWaterSens. We select Northern Indiana in the United States as the primary area of study with a focus of nitrate contamination. Nitrate (NO_3^--N) is a critical water contaminant that can lead to increased risk of certain cancers, birth defects, and thyroid problems resulting from prolonged exposure [3]. Notably, Northern Indiana is a high-risk area of nitrate contamination due to the high application rate of nitrogen fertilizer in agricultural and residential land use [49]. More importantly, a large number of households in Northern Indiana are well-dependent and at high risk of elevated nitrate concentration in their drinking water [49]. While nitrate is the primary groundwater contaminant investigated in this study, the presented crowdsensing data collection approach can be easily adapted to collect crowdsensing groundwater contamination data for other water contaminants (e.g., atrazine, phosphorus) with proper contaminant test kits and generalized to other geographic areas with high-risk of groundwater contamination.

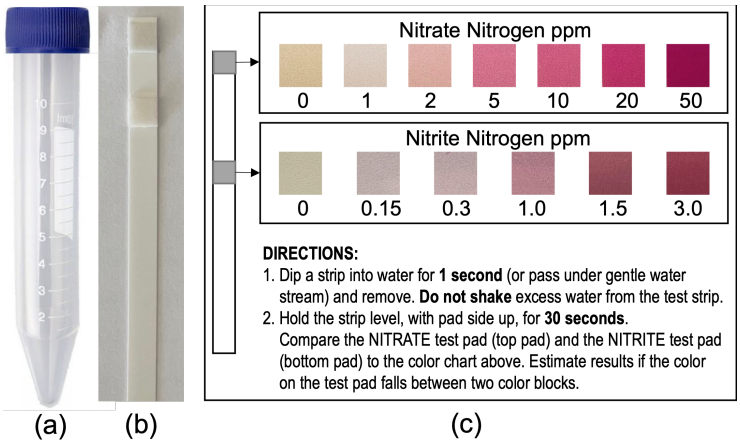


Image (a) is the test tube used for test strip dipping and tap water sample collection. **Image (b)** is the test strip that is used for testing the nitrate concentration in the water sample of interest. **Image (c)** is the colorimetric chart for the participants to determine the nitrate concentration results on the dipped test strips. Please note that the unit of the nitrate concentration measured by the colorimetric chart is *parts per million (ppm)* which is equivalent to mg/L (i.e., 1 ppm = 1 mg/L) for measuring contaminant in water.

Figure 4: Example of Test Kit for Nitrate Concentration

5.1. Participant Recruitment and Nitrate Test Kit

The crowdsensing participants were recruited from well-dependent residential communities in St. Joseph County, Northern Indiana, through the outreach program at the University of Notre Dame. We strictly followed the corresponding IRB protocol for participant recruitment and data collection in this study. Proper participant consents were obtained prior to the experiments. In our experiments, we distribute the nitrate test kits to the participants for measuring the crowdsensing nitrate concentration. We show an example of the nitrate test kit used in our study in Figure 4. In particular, during the participant recruitment process, the participants are instructed on how to use the test kit to: 1) collect tap water samples in a water tube (Figure 4 (a)), and 2) measure the crowdsensing nitrate concentration of water samples using the test strip (Figure 4 (b)) and colorimetric chart (Figure 4 (c)).

5.2. Crowdsensing Data Collection

In our experiments, we perform two case studies to collect the crowdsensing nitrate contamination data under different experimental settings. In particular, the first case study (Case Study I) is designed to collect crowdsensing nitrate concentration measurements *with* the groundwater samples to be used for validating the ground-truth nitrate concentration. However, during the global COVID-19 pandemic, it is impractical to collect groundwater samples from the participants due to various societal restrictions (e.g., stay-at-home orders, quarantines) and the participants' health concerns. In light of such a challenge, we further design a second case study (Case Study II) to collect the crowdsensing nitrate concentration data in a fully at-home setting *without* requiring the participants to bring in the groundwater samples for ground-truth nitrate concentration validation. We elaborate on the two case studies in detail below.

5.2.1. Case Study I (Before COVID-19 Pandemic)

The first case study was conducted in the Fall of 2019 for a data collection period of four weeks. An overview of the data collection pipeline is shown in Figure 5. In particular, we asked the participants to measure the nitrate concentration of the tap water in their households 3 times a week (i.e., on Monday, Wednesday, and Friday). The crowdsensing nitrate concentration was measured by dipping the test strip with the tap water from the participant's household and comparing the dipped test strip with the colorimetric chart to obtain the approximate measurement of nitrate concentration (Figure 4). The measured nitrate concentrations were recorded via online and/or paper form depending on each participant's accessibility to the Internet. The participants were also asked to report the anthropogenic context (Definition 6) along with their measurements. Finally, the participants returned the collected tap water samples, which were collected and transported to the university laboratory for the professional nitrate concentration measurement. In our study, we used the cadmium reduction method on a Lachat QuikChem™Autoanalyzer to measure the ground-truthed nitrate concentration [50]. To ensure the integrity and quality of the ground-truth data, we excluded invalid water samples that were not properly stored in the sample tube or have sample tube labels that could not be matched with the crowdsensing nitrate measurements for

ground truth validation.

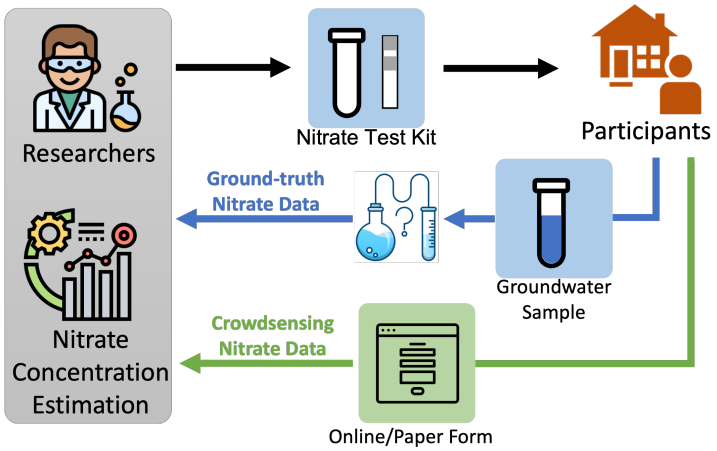


Figure 5: Crowdsensing Data Collection Pipeline - Case Study I

We totally received 161 weekly crowdsensing data samples with ground-truth nitrate concentration, where each weekly crowdsensing data sample contained 3 crowdsensing measurements. The size of the crowdsensing dataset matches the scale of the datasets studied in state-of-the-art groundwater contamination estimation literature that is shown to be sufficient to evaluate groundwater nitrate concentration estimation models for a regular watershed [51]. A summary of the collected crowdsensing and ground-truth data is presented in Table 1. We observe that the average crowdsensing nitrate concentration is lower than the average ground-truth nitrate concentration. A possible reason is that the crowd participants may underestimate the test strip readings when the color of the dipped test strips falls between two color blocks. The crowdsensing measurement error (i.e., the difference between the crowdsensing and ground-truth nitrate concentration) is particularly significant when the ground-truth nitrate concentration is high due to the low colorimetric precision of the test strips for high nitrate concentration (e.g., above 2 mg/L).

5.2.2. Case Study II (During COVID-19 Pandemic)

The second case study was conducted in the Fall of 2020 and the Spring of 2021 with an adapted data collection procedure to accommodate the restrictions during the

Data Trace	Crowdsensing	Ground Truth
Mean	0.78	1.10
Minimum	0.00	0.00
Maximum	8.00	9.33
Standard Deviation	1.21	1.75

Table 1: Summary Statistics of Crowdsensing Nitrate Data (mg/L) - Case Study I

COVID-19 pandemic. We showed an overview of the crowdsensing data collection pipeline in Figure 6. The goal of the adaptation was to eliminate the requirement in the original experimental design (i.e., Case Study I) that asked the participants to collect and bring in the groundwater samples, which was often infeasible given various societal restrictions during the COVID-19 pandemic. In particular, we modified the nitrate test kit in Case Study I (Section 5.2.1) by including three pairs of test strip and nitrate water solution sample (Figure 6) at pre-identified ground-truth nitrate concentrations that belong to the following three levels: low (0-1 mg/L), medium (1-10 mg/L) and high (10-20 mg/L). We distributed the test kits to our participants without disclosing the ground-truth nitrate concentration of the nitrate solutions in the test kits. For each pair of the test strip and nitrate water sample, we asked the participants to measure the nitrate concentration using the test strip (Figure 4(b)) and compare the dipped test strip with the colorimetric (Figure 4(c)). The participants were asked to record their crowdsensing measurements and submit them through an online submission system. Such a process is repeated for all three nitrate water samples in the test kits.

We totally received 177 crowdsensing data samples where each data sample contains the crowdsensing measurements of three NO_3^--N solutions at different pre-identified ground-truth nitrate concentration levels. A summary of the collected crowdsensing and ground-truth data is presented in Table 2. We observe similar patterns as in Case Study I where the crowdsensing measurement errors are higher for measuring the nitrate solution with high nitrate concentration resulting from the low colorimetric precision of the test strips at high nitrate concentration levels.

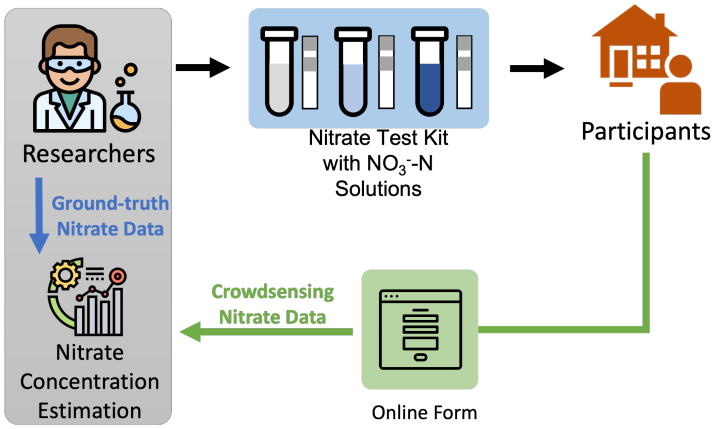


Figure 6: Crowdsensing Data Collection Pipeline - Case Study II

Data Trace	Nitrate Concentration Level		
	Low (0-1 mg/L)	Medium (1-10 mg/L)	High (10-20 mg/L)
Mean	1.02	5.51	16.55
Minimum	0.00	1.00	0.00
Maximum	5.00	35.00	50.00
Standard Deviation	0.93	4.48	13.70

Table 2: Summary Statistics of Crowdsensing Nitrate Data (mg/L) - Case Study II

6. Evaluation

In this section, we evaluate the nitrate concentration estimation performance of the proposed CrowdWaterSens framework using the crowdsensing datasets collected from the two case studies as described in Section 5. Evaluation results demonstrate that CrowdWaterSens achieves substantial performance gains compared to the SmartWaterSens we developed in our conference paper and the state-of-the-art groundwater contamination estimation solutions in accurately estimating nitrate concentration in both case studies.

6.1. Baseline Methods

We compare CrowdWaterSens with a set of state-of-the-art methods for groundwater nitrate concentration estimation.

- **SmartWaterSens**: A context-aware graph neural network framework that is developed in our conference paper [17]. SmartWaterSens explores the anthropogenic context of crowdsensing nitrate data to estimate the nitrate concentration in crowdsensing groundwater samples.

- **NCE**: A naive crowd estimation baseline method that directly uses the averaged crowdsensing measurements of nitrate concentration as the estimated nitrate concentration.

- **ELR**: It is an ensemble learning based regression method that utilizes spatial hydrogeological features to predict nitrate concentration in groundwater [22]. We adapted ELR to estimate nitrate concentration using the crowdsensing nitrate concentration and location context information.

- **NCRA**: A nitrate contamination risk assessment solution that leverages the sampled nitrate concentration (i.e., nitrate measurement from sampled wells) to estimate the groundwater nitrate pollution in nearby locations [52]. We replace the sampled nitrate concentration with the crowdsensing nitrate concentration obtained from the crowd participants.

- **GWR**: A geographically weighted regression scheme that learns the weighted regression coefficients based on the distance between the sampling site locations and the target location at which the nitrate concentration is estimated [7]. In particular, we use the pairwise Haversine distance between the crowd locations in the training and testing sets as the distance weights in GWR.

- **AVI**: A deep learning water contamination estimation approach that builds a multilayer perceptron neural network to estimate the concentration of groundwater nitrate [53]. In particular, we use the crowdsensing measurements and the anthropogenic context features as the input to AVI and predict the estimated nitrate concentration.

- **BNN**: A Bayesian neural network framework that utilizes Bayesian optimization to estimate prediction uncertainty and improve the nitrate estimation performance [51]. Specifically, BNN takes the crowdsensing measurements and the

anthropogenic context information as input features to estimate nitrate concentration.

- **GCN**: A principled spatio-temporal graph convolutional neural network solution that jointly aggregates the spatial and temporal relations of graph nodes for traffic flow prediction [54]. We adapted GCN with the crowdsensing contamination network (CCN) constructed in our CrowdWaterSens to estimate the nitrate concentration of each node in CCN.

- **GTN**: A generalized graph transformer network that fuses propositional node features in graph neural networks for regression tasks [55]. We train GTN using the crowdsensing nitrate data to predict the concentration of groundwater nitrate in our study.

6.2. Experiment Settings

To ensure a fair comparison, we use the same input features to train CrowdWaterSens and all compared baselines, except the NCE baseline, which only uses the crowdsensing measurements as the estimated nitrate concentration. In our experiments, we use 80% of the dataset as a *training-validation* set and the remaining 20% as a *test* set. We perform 5-fold cross-validation on the training-validation set to tune hyperparameters of all compared methods and evaluate the estimation performance based on the test set. In addition, for the experiments in Case Study I, we use the mean crowdsensing measurements in each week as the input crowdsensing measurements (i.e., c_i in Equation 2) collected from the participants to estimate the true nitrate concentration in the groundwater samples, and use the variance of the crowdsensing measurements in each week as the crowd sensor uncertainty in the UNO module. For the experiments in Case Study II, we focus on the medium-level nitrate solution in our study and use the normalized estimation error on the low-level and high-level nitrate solutions to compute the crowd sensor uncertainty in the UNO module. This is because nitrate concentration in the medium level (i.e., 1-10 mg/L) is the most concerned range of groundwater nitrate concentration in well-dependent communities where nitrate concentrations

greater than 3 mg/L generally indicate contamination in groundwater.[3] In addition, we observe that nitrate concentration in the high level (i.e., 10-20 mg/L) rarely happens in our studied well-dependent communities according to the nitrate concentration measured in the groundwater samples collected in Case Study I (Table 1). Therefore, we adopt the crowdsensing measurements on the medium level as our evaluation dataset to study the performance of CrowdWaterSens on groundwater nitrate concentration in well-dependent communities.

We implement our CrowdWaterSens model using PyTorch 1.10[4] and run our experiments on Ubuntu 20.04 with four NVIDIA A40. We set the total number of epochs as 50 and train CrowdWaterSens at an initial learning rate of 0.001 with a decay of 0.95 in each epoch. We adopt a set of evaluation metrics that are commonly used for evaluating regression models. In particular, let $\hat{y}_i$ and y_i be the *estimated* and *ground-truth* nitrate concentration of the i^{th} sample in a test set of N samples, we consider the following evaluation metrics:

- **Mean Absolute Error (MAE):** MAE $= \frac{1}{N}\sum_{i=1}^{N}|\hat{y}_i - y_i|$
- **Mean Squared Error (MSE):** MSE $= \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2$
- **Root Mean Squared Error (RMSE):** RMSE $= \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}$
- **Coefficient of Determination (R^2):**
  $R^2 = 1 - [\sum_{i=1}^{N}(y_i - \hat{y}_i)^2]/[\sum_{i=1}^{N}(y_i - \bar{y})^2]$ where $\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$

Intuitively, a lower value of MAE, MSE, and RMSE, and a higher value of R^2 represent a more accurate nitrate concentration estimation.

6.3. Nitrate Concentration Estimation Performance

In the first set of experiments, we evaluate the performance of nitrate concentration estimation using the crowdsensing nitrate datasets collected from the two case studies. We report the evaluation results on Case Study I and Case Study II in Table 3

[3] https://www.epa.gov/nutrient-policy-data/estimated-nitrate-concentrations-groundwater-used-drinking

[4] https://pytorch.org/

and Table 4, respectively. We observe that our CrowdWaterSens achieves substantial performance improvements compared to all baseline methods in terms of all evaluation metrics on both case studies. In particular, CrowdWaterSens outperforms the best-performing baseline (i.e., SmartWaterSens) by 7.1%, 4.3%, 4.1%, and 10.2%, in terms of MAE, MSE, RMSE, and R^2 in Case Study I, respectively. We observe similar performance gains in Case Study II as well. However, we also observe that the nitrate estimation errors of all methods in Case Study II are higher than the ones in Case Study I. This is mainly due to the difference of the nitrate concentration in the water samples between the two case studies, where the ground-truth nitrate concentrations in Case Study I are in general much lower than the ground-truth nitrate concentrations in Case Study II (Table 1 and 2). Such a difference also amplifies the crowdsensing nitrate measurement errors given the fact that crowdsensing measurements are less accurate on nitrate concentration above 2 mg/L due to the limited colorimetric precision of the test strips. Such significant performance improvements demonstrate the effectiveness of CrowdWaterSens in accurately estimating the groundwater nitrate contamination through the novel spatio-temporal crowdsensing contamination network. In addition, we also attribute the performance gains of CrowdWaterSens to the principled uncertainty network optimization strategy that explicitly examines the uncertainty of crowdsensing nitrate measurements from unreliable crowd sensors to optimize the nitrate estimation performance in crowdsensing-based groundwater monitoring systems.

6.4. Ablation Study

In the second set of experiments, we conduct an ablation study to investigate the contribution and effectiveness of major components in the CrowdWaterSens framework. In particular, we consider the following model variants of CrowdWaterSens.

- **CrowdWaterSens\S**: The variant of CrowdWaterSens that ignores the spatial distance between each pair of the crowdsensing nodes in CCN and assigns the same weight to the spatial edges.
- **CrowdWaterSens\T**: The variant of CrowdWaterSens that excludes the temporal relations between the crowdsensing nodes in different sensing cycles by

Method	MAE	MSE	RMSE	R^2
CrowdWaterSens	**0.4639**	**0.4817**	**0.6829**	**0.3618**
SmartWaterSens	0.4993	0.5032	0.7125	0.3247
NCE	1.0327	2.9376	1.7783	-0.2543
ELR	0.7794	1.6832	1.2784	-0.1407
NCRA	0.5376	0.6597	0.8133	0.2970
GWR	0.7926	1.1247	1.1359	0.0823
AVI	0.7326	0.7548	0.8837	-0.1032
BNN	0.7821	0.8733	0.9417	-0.2015
GCN	0.6849	0.8446	0.8973	0.0372
GTN	0.8734	0.9728	1.1873	-0.1769

Table 3: Nitrate Concentration Estimation Performance - Case Study I

Method	MAE	MSE	RMSE	R^2
CrowdWaterSens	**0.8932**	**2.9836**	**1.7729**	**0.3374**
SmartWaterSens	0.9628	3.1726	1.8932	0.3089
NCE	1.9722	11.0278	3.3208	-0.0459
ELR	1.2376	6.7738	2.3849	0.1839
NCRA	1.0287	4.0215	2.1634	0.2337
GWR	1.3486	7.3849	2.5381	0.1329
AVI	1.4295	7.9447	2.8328	0.0916
BNN	1.4872	8.2476	3.0182	-0.1837
GCN	1.2293	6.2145	2.2935	0.1748
GTN	1.1834	5.6724	2.2403	0.1183

Table 4: Nitrate Concentration Estimation Performance - Case Study II

removing the temporal edges in CCN.

- **CrowdWaterSens\C**: The variant of CrowdWaterSens that excludes the anthropogenic context of the crowdsensing nodes and only uses the crowdsensing measurements as the node feature in DR-GCN.

- **CrowdWaterSens\U**: The variant of CrowdWaterSens that removes the uncertainty attention layer in DR-GCN and considers each node as equally important in CCN.

We summarize the evaluation results of the ablation study on the datasets collected from Case Study I and II in Table 5 and Table 6, respectively. Please note that our Case Study II is designed to collect the crowdsensing data in a single round (i.e., the participants were asked to only measure the nitrate concentration once on multiple nitrate solutions) and does not capture the temporal dynamics. Therefore, the results of CrowdWaterSens\T are the same as CrowdWaterSens on the dataset from Case Study II and thus are omitted in Table 6. We observe that CrowdWaterSens achieves the best performance when it integrates all components in the framework. The evaluation results demonstrate the effectiveness and necessity of the key components in CrowdWaterSens.

Method	MAE	MSE	RMSE	R^2
CrowdWaterSens	**0.4639**	**0.4817**	**0.6829**	**0.3618**
CrowdWaterSens\S	0.5327	0.5287	0.7748	0.1938
CrowdWaterSens\T	0.5106	0.5172	0.7183	0.2174
CrowdWaterSens\C	0.4985	0.5013	0.7031	0.2635
CrowdWaterSens\U	0.5182	0.5216	0.7227	0.2218

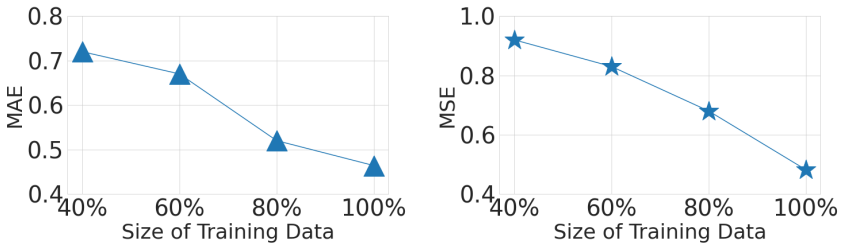
Table 5: Ablation Study - Case Study I

Method	MAE	MSE	RMSE	R^2
CrowdWaterSens	**0.8932**	**2.9836**	**1.7729**	**0.3374**
CrowdWaterSens\S	0.9026	3.1195	1.8278	0.3185
CrowdWaterSens\C	0.8983	3.0837	1.8026	0.3027
CrowdWaterSens\U	1.1277	6.0739	2.2334	0.1039

Table 6: Ablation Study - Case Study II

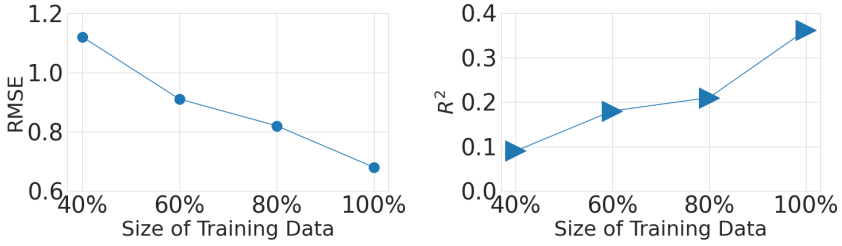
6.5. *Effect of Training Data Size*

In the third set of experiments, we further investigate the effect of training data size on the performance of nitrate concentration estimation in CrowdWaterSens. In particular, we vary the size of training data from 40% to 100% of the entire training data and report the nitrate concentration estimation performance on the testing set in Figure 7 and Figure 8 for Case Study I and II, respectively. We observe that the nitrate estimation performance consistently increases as we increase the amount of training data, indicating that additional training data will have great potential to enhance the model performance. Such observation also encourages our future work to further extend the number of crowdsensing communities and recruit additional crowdsourcing participants in our study.



(a) Estimation Performance (MAE)

(b) Estimation Performance (MSE)



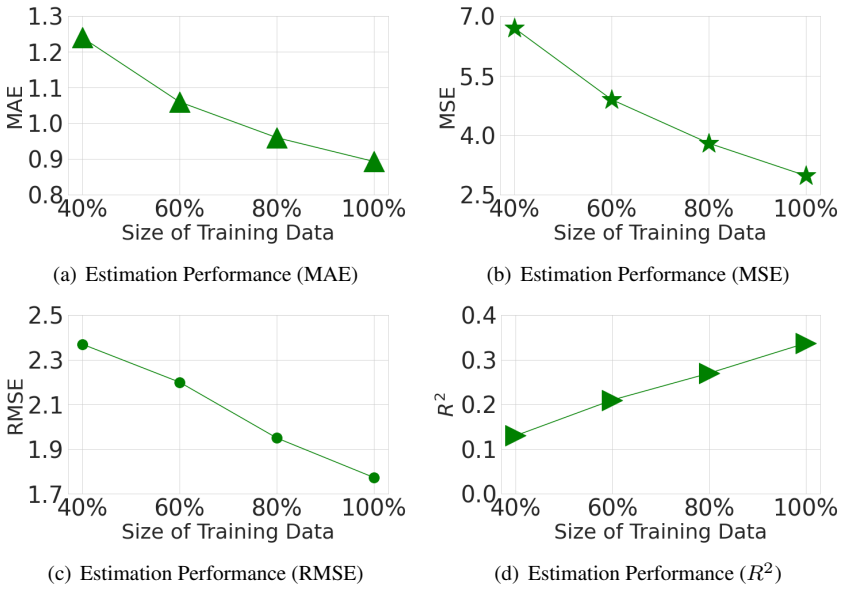
(c) Estimation Performance (RMSE)

(d) Estimation Performance (R^2)

Figure 7: Estimation Performance vs. Training Data Size (Case Study I)

7. Discussion

In this study, we focus on the crowdsensing-based groundwater contamination estimation problem for groundwater quality monitoring in well-dependent communities in the United States. We note that generalizability is an important aspect and therefore



(a) Estimation Performance (MAE)

(b) Estimation Performance (MSE)

(c) Estimation Performance (RMSE)

(d) Estimation Performance (R^2)

Figure 8: Estimation Performance vs. Training Data Size (Case Study II)

we highlight a few criteria for applying the CrowdWaterSens framework to monitor groundwater quality in different scenarios (e.g., different geographical areas, groundwater contaminants). First, CrowdWaterSens is designed to estimate groundwater contamination in suburban and rural communities which do not have central water supply systems and rely on groundwater from private wells as the primary drinking water resource. There are a number of countries with well-dependent communities, including the United States [56], Canada [57], Brazil [58], and Italy [59], to name a few. Second, the CrowdWaterSens may be applied in the areas with groundwater contaminants that can be monitored by affordable test strips (e.g., lead [60], chlorine [61]). Third, a reasonable number of participants need to be recruited, trained, and engaged for the crowdsensing experiments to ensure the estimation accuracy of the contaminant concentration. In the case studies, we collaborated with local school systems to demonstrate the feasibility of recruiting participants for crowdsensing-based groundwater monitoring. We believe such a strategy can be generalized to recruit crowdsensing participants in common well-dependent communities to deploy the CrowdWaterSens framework. In all, if the above criteria are satisfied, we envision the Crowd-

WaterSens framework has great potential to be generalized to a broader range of geographical regions (e.g., countries beyond the U.S.) to monitor groundwater quality in well-dependent communities.

We also acknowledge that there may exist certain scenarios that do not meet the above criteria. For example, an area may rely on the central water supply as the primary water resource. Then, it may be unnecessary to apply the CrowdWaterSens framework to monitor the groundwater quality because central water supply systems are routinely monitored in the water supplying utilities. Alternatively, our CrowdWaterSens framework could complement existing water quality management systems by monitoring and estimating the water contamination occurring in water distribution systems (e.g., lead contamination caused by pipe corrosion [62]) at the household end. Therefore, we believe there are many opportunities for CrowdWaterSens to be applied beyond the aforementioned criteria and will continue to explore the potential opportunities in our future work.

Moreover, we note that the crowdsensing participants have played an essential role in the CrowdWaterSens framework. In this work, we recruited and trained our crowdsensing participants through the local high school systems to collect the crowdsensing data for the experiments. While such a strategy is applicable to common residential communities with municipal government, it may limit the participant population to a determined group of high school students who have similar backgrounds (e.g., age, education). Such a group of participants may have limited knowledge about the anthropogenic context of their households (e.g., farm proximity, fertilizer application), leading to inaccurate crowdsensing data. To overcome such a limitation, we plan to further explore the role of participants in CrowdWaterSens by expanding our participant population to community residents beyond high school students. A possible solution is to leverage the local homeowner associations or direct mail services to reach out to a larger group of residents in the community of interest. In the meanwhile, we also plan to design an anonymized survey to obtain the participant background information associated with the crowdsensing measurements, which can be further leveraged to enhance the estimation accuracy of CrowdWaterSens.

8. Conclusion

This paper presents CrowdWaterSens, a crowdsensing-based scheme to address the groundwater nitrate concentration estimation problem in well-dependent communities. CrowdWaterSens designs a spatio-temporal crowdsensing contamination network to explicitly model the underlying correlation between crowdsensing nitrate measurements and the anthropogenic context of the studied areas to accurately estimate the groundwater nitrate concentration. We carry out two real-world case studies using crowdsensing nitrate measurements collected from Northern Indiana, United States, to evaluate the CrowdWaterSens framework. Evaluation results show that CrowdWaterSens significantly outperforms state-of-the-art contamination estimation solutions in accurately estimating nitrate concentration in groundwater.

Acknowledgement

We thank the crowd participants in well-dependent communities for their participation. This research is supported in part by the National Science Foundation under Grant No. CHE-2105032, IIS-2008228, CNS-1845639, CNS-1831669. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

References

[1] P. Li, D. Karunanidhi, T. Subramani, K. Srinivasamoorthy, Sources and consequences of groundwater contamination, Archives of environmental contamination and toxicology 80 (1) (2021) 1–10.

[2] F. Kon, K. Braghetto, E. Z. Santana, R. Speicys, J. G. Guerra, Toward smart and sustainable cities, Communications of the ACM 63 (11) (2020) 51–52.

[3] A. Rahman, N. Mondal, K. Tiwari, Anthropogenic nitrate in groundwater and its health risks in the view of background concentration in a semi arid area of rajasthan, india, Scientific reports 11 (1) (2021) 1–13.

[4] S. L. Speir, L. Shang, D. Bolster, J. L. Tank, C. J. Stoffel, D. M. Wood, B. W. Peters, N. Wei, D. Wang, Solutions to current challenges in widespread monitoring of groundwater quality via crowdsensing, Groundwater 60 (1) (2022) 15–24.

[5] R. Weinmeyer, A. Norling, M. Kawarski, E. Higgins, The safe drinking water act of 1974 and its role in providing access to safe drinking water in the united states, AMA Journal of Ethics 19 (10) (2017) 1018–1026.

[6] D. Machiwal, M. K. Jha, V. P. Singh, C. Mohan, Assessment and mapping of groundwater vulnerability to pollution: Current status and challenges, Earth-Science Reviews 185 (2018) 901–927.

[7] E.-H. Koh, E. Lee, K.-K. Lee, Application of geographically weighted regression models to predict spatial characteristics of nitrate contamination: Implications for an effective groundwater management strategy, Journal of Environmental Management 268 (2020).

[8] L. Knoll, L. Breuer, M. Bach, Nation-wide estimation of groundwater redox conditions and nitrate concentrations through machine learning, Environmental Research Letters 15 (6) (2020).

[9] Z. He, P. Zhang, L. Wu, A. M. Rocha, Q. Tu, Z. Shi, B. Wu, Y. Qin, J. Wang, Q. Yan, et al., Microbial functional gene diversity predicts groundwater contamination and ecosystem functioning, MBio 9 (1) (2018) e02435–17.

[10] D. Wang, M. T. Amin, S. Li, T. Abdelzaher, L. Kaplan, S. Gu, C. Pan, H. Liu, C. C. Aggarwal, R. Ganti, et al., Using humans as sensors: an estimation-theoretic perspective, in: IPSN-14 proceedings of the 13th international symposium on information processing in sensor networks, IEEE, 2014, pp. 35–46.

[11] D. Wang, T. Abdelzaher, L. Kaplan, C. C. Aggarwal, Recursive fact-finding: A streaming approach to truth estimation in crowdsourcing applications, in: 2013 IEEE 33rd international conference on distributed computing systems, IEEE, 2013, pp. 530–539.

[12] A. L. Srivastav, Chemical fertilizers and pesticides: role in groundwater contamination, in: Agrochemicals detection, treatment and remediation, Elsevier, 2020, pp. 143–159.

[13] Y. Zhang, R. Zong, Z. Kou, L. Shang, D. Wang, A crowd-driven dynamic neural architecture searching approach to quality-aware streaming disaster damage assessment, in: 2021 IEEE/ACM 29th International Symposium on Quality of Service (IWQOS), IEEE, 2021, pp. 1–6.

[14] S. Liu, Z. Zheng, F. Wu, S. Tang, G. Chen, Context-aware data quality estimation in mobile crowdsensing, in: IEEE INFOCOM 2017-IEEE Conference on Computer Communications, IEEE, 2017, pp. 1–9.

[15] M. Spurling, C. Hu, H. Zhan, V. S. Sheng, Estimating crowd-worker's reliability with interval-valued labels to improve the quality of crowdsourced work, in: 2021 IEEE Symposium Series on Computational Intelligence (SSCI), IEEE, 2021, pp. 01–08.

[16] H. Lan, Y. Pan, A crowdsourcing quality prediction model based on random forests, in: 2019 IEEE/ACIS 18th International Conference on Computer and Information Science (ICIS), IEEE, 2019, pp. 315–319.

[17] L. Shang, Y. Zhang, Q. Ye, N. Wei, D. Wang, Smartwatersens: A crowdsensing-based approach to groundwater contamination estimation, in: 2022 IEEE International Conference on Smart Computing (SMARTCOMP), IEEE, 2022, pp. 48–55.

[18] Y. Wu, Y. Chen, N. Wei, Biocatalytic properties of cell surface display laccase for degradation of emerging contaminant acetaminophen in water reclamation, Biotechnology and bioengineering 117 (2) (2020) 342–353.

[19] X. Long, F. Liu, X. Zhou, J. Pi, W. Yin, F. Li, S. Huang, F. Ma, Estimation of spatial distribution and health risk by arsenic and heavy metals in shallow groundwater around dongting lake plain using gis mapping, Chemosphere 269 (2021) 128698.

[20] P. Li, S. He, N. Yang, G. Xiang, Groundwater quality assessment for domestic and agricultural purposes in yan'an city, northwest china: implications to sustainable groundwater quality management on the loess plateau, Environmental Earth Sciences 77 (23) (2018) 1–16.

[21] J. C. Egbueri, Groundwater quality assessment using pollution index of groundwater (pig), ecological risk index (eri) and hierarchical cluster analysis (hca): a case study, Groundwater for Sustainable Development 10 (2020) 100292.

[22] L. Knoll, L. Breuer, M. Bach, Large scale prediction of groundwater nitrate concentrations from spatial data using machine learning, Science of the total environment 668 (2019) 1317–1327.

[23] D. Wang, T. Abdelzaher, L. Kaplan, Social sensing: building reliable systems on unreliable data, Morgan Kaufmann, 2015.

[24] Y. Zhang, X. Dong, D. Zhang, D. Wang, A syntax-based learning approach to geo-locating abnormal traffic events using social sensing, in: 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), IEEE, 2019, pp. 663–670.

[25] Y. Zhang, R. Zong, L. Shang, Z. Kou, D. Wang, A deep contrastive learning approach to extremely-sparse disaster damage assessment in social sensing, in: Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, 2021, pp. 151–158.

[26] Y. Zhang, R. Zong, L. Shang, M. T. Rashid, D. Wang, Superclass: A deep duo-task learning approach to improving qos in image-driven smart urban sensing applications, in: 2021 IEEE/ACM 29th International Symposium on Quality of Service (IWQOS), IEEE, 2021, pp. 1–6.

[27] L. Shang, Y. Zhang, C. Youn, D. Wang, Sat-geo: A social sensing based content-only approach to geolocating abnormal traffic events using syntax-based probabilistic learning, Information Processing & Management 59 (2) (2022) 102807.

[28] Y. Liang, D. Sengupta, M. J. Campmier, D. M. Lunderberg, J. S. Apte, A. H. Goldstein, Wildfire smoke impacts on indoor air quality assessed using crowdsourced data in california, Proceedings of the National Academy of Sciences 118 (36) (2021) e2106478118.

[29] M. Silva, G. Signoretti, J. Oliveira, I. Silva, D. G. Costa, A crowdsensing platform for monitoring of vehicular emissions: A smart city perspective, Future Internet 11 (1) (2019) 13.

[30] Y. Zhang, Y. Lu, D. Zhang, L. Shang, D. Wang, Risksens: A multi-view learning approach to identifying risky traffic locations in intelligent transportation systems using social and remote sensing, in: 2018 IEEE International Conference on Big Data (Big Data), IEEE, 2018, pp. 1544–1553.

[31] L. Breuer, N. Hiery, P. Kraft, M. Bach, A. H. Aubert, H.-G. Frede, Hydrocrowd: a citizen science snapshot to assess the spatial control of nitrogen solutes in surface waters, Scientific reports 5 (1) (2015) 1–10.

[32] H. P. Lee, S. Garg, K. M. Lim, Crowdsourcing of environmental noise map using calibrated smartphones, Applied Acoustics 160 (2020) 107130.

[33] S. Wang, J. Cao, P. Yu, Deep learning for spatio-temporal data mining: A survey, IEEE transactions on knowledge and data engineering (2020).

[34] D. Pavlyuk, Feature selection and extraction in spatiotemporal traffic forecasting: a systematic literature review, European Transport Research Review 11 (1) (2019) 1–19.

[35] H. Werneck, N. Silva, M. C. Viana, F. Mourão, A. C. Pereira, L. Rocha, A survey on point-of-interest recommendation in location-based social networks, in: Proceedings of the Brazilian Symposium on Multimedia and the Web, 2020, pp. 185–192.

[36] X. Ren, X. Li, K. Ren, J. Song, Z. Xu, K. Deng, X. Wang, Deep learning-based weather prediction: a survey, Big Data Research 23 (2021) 100178.

[37] X. Luo, D. Li, Y. Yang, S. Zhang, Spatiotemporal traffic flow prediction with knn and lstm, Journal of Advanced Transportation 2019 (2019).

[38] C. Liu, J. Liu, S. Xu, J. Wang, C. Liu, T. Chen, T. Jiang, A spatiotemporal dilated convolutional generative network for point-of-interest recommendation, ISPRS International Journal of Geo-Information 9 (2) (2020) 113.

[39] R. Castro, Y. M. Souto, E. Ogasawara, F. Porto, E. Bezerra, Stconvs2s: Spatiotemporal convolutional sequence to sequence network for weather forecasting, Neurocomputing 426 (2021) 285–298.

[40] V. S. Sheng, J. Zhang, Machine learning with crowdsourcing: A brief summary of the past research and future directions, in: Proceedings of the AAAI conference on artificial intelligence, Vol. 33, 2019, pp. 9837–9843.

[41] F. Brühlmann, S. Petralito, L. F. Aeschbach, K. Opwis, The quality of data collected online: An investigation of careless responding in a crowdsourced sample, Methods in Psychology 2 (2020) 100022.

[42] N. Papapesios, C. Ellul, A. Shakir, G. Hart, Exploring the use of crowdsourced geographic information in defence: challenges and opportunities, Journal of Geographical Systems 21 (1) (2019) 133–160.

[43] X. Gong, N. Shroff, Incentivizing truthful data quality for quality-aware mobile data crowdsourcing, in: Proceedings of the Eighteenth ACM International Symposium on Mobile Ad Hoc Networking and Computing, 2018, pp. 161–170.

[44] A. F. Probert, D. Wegmann, L. Volery, T. Adriaens, R. Bakiu, S. Bertolino, F. Essl, E. Gervasini, Q. Groom, G. Latombe, et al., Identifying, reducing, and communicating uncertainty in community science: a focus on alien species, Biological Invasions (2022) 1–27.

[45] D. Wang, L. Kaplan, H. Le, T. Abdelzaher, On truth discovery in social sensing: A maximum likelihood estimation approach, in: Proceedings of the 11th international conference on Information Processing in Sensor Networks, 2012, pp. 233–244.

[46] W. A. Sigler, S. A. Ewing, C. A. Jones, R. A. Payn, P. Miller, M. Maneta, Water and nitrate loss from dryland agricultural soils is controlled by management, soils, and weather, Agriculture, Ecosystems & Environment 304 (2020) 107158.

[47] C. C. Robusto, The cosine-haversine formula, The American Mathematical Monthly 64 (1) (1957) 38–40.

[48] M. Schlichtkrull, T. N. Kipf, P. Bloem, R. v. d. Berg, I. Titov, M. Welling, Modeling relational data with graph convolutional networks, in: European semantic web conference, Springer, 2018.

[49] I. D. of Environmental Management, Statewide groundwater monitoring network: Summary and results (2016).

[50] A. P. H. Association, A. W. W. Association, W. E. Federation, et al., Standard methods for the examination of water and wastewater, American Public Health Association, 2017.

[51] S. S. Band, S. Janizadeh, S. C. Pal, I. Chowdhuri, Z. Siabi, A. Norouzi, A. M. Melesse, M. Shokri, A. Mosavi, Comparative analysis of artificial intelligence models for accurate estimation of groundwater nitrate concentration, Sensors 20 (20) (2020) 5763.

[52] F. Sajedi-Hosseini, A. Malekian, B. Choubin, O. Rahmati, S. Cipullo, F. Coulon, B. Pradhan, A novel machine learning-based approach for the risk assessment of nitrate groundwater contamination, Science of the total environment 644 (2018) 954–962.

[53] H. E. Elzain, S. Y. Chung, V. Senapathi, S. Sekar, N. Park, A. A. Mahmoud, Modeling of aquifer vulnerability index using deep learning neural networks coupling with optimization algorithms, Environmental Science and Pollution Research 28 (40) (2021).

[54] A. Ali, Y. Zhu, M. Zakarya, Exploiting dynamic spatio-temporal graph convolutional neural networks for citywide traffic flows prediction, Neural networks 145 (2022) 233–247.

[55] V. P. Dwivedi, X. Bresson, A generalization of transformer networks to graphs, arXiv preprint arXiv:2012.09699 (2020).

[56] F. Edition, Guidelines for drinking-water quality, WHO chronicle 38 (4) (2011) 104–8.

[57] P. D. Hynds, M. K. Thomas, K. D. M. Pintar, Contamination of groundwater systems in the us and canada by enteric pathogens, 1990–2013: a review and pooled-analysis, PloS one 9 (5) (2014) e93301.

[58] R. Hirata, F. Cagnon, A. Bernice, C. H. Maldaner, P. Galvão, C. Marques, R. Terada, C. Varnier, M. C. Ryan, R. Bertolo, Nitrate contamination in brazilian urban aquifers: A tenacious problem, Water 12 (10) (2020) 2709.

[59] E. Abascal, L. Gómez-Coma, I. Ortiz, A. Ortiz, Global diagnosis of nitrate pollution in groundwater and review of removal technologies, Science of the total environment 810 (2022) 152233.

[60] Z. Wang, B. Chen, J. Duan, T. Hao, X. Jiang, Z. Guo, S. Wang, A test strip for lead (ii) based on gold nanoparticles multi-functionalized by dnazyme and barcode dna., Journal of Analytical Chemistry 70 (3) (2015).

[61] C. Huangfu, Y. Zhang, M. Jang, L. Feng, A μpad for simultaneous monitoring of cu2+, fe2+ and free chlorine in drinking water, Sensors and Actuators B: Chemical 293 (2019) 350–356.

[62] P. Levallois, P. Barn, M. Valcke, D. Gauvin, T. Kosatsky, Public health consequences of lead in drinking water, Current environmental health reports 5 (2018) 255–262.