A Provenance Scheme for Emerging Water Contaminants

Dian Yang[1], Yongxin Zhu[1], Tian Huang[1], Yiliang He[2], Stéphane Bressan[3]

[1] *School of Microelectronics, Shanghai Jiao Tong University, Shanghai, China*

[2] *School of Environmental Science & Engineering, Jiao Tong University Shanghai, China*

[3]*School of Computing, National University of Singapore, Singapore*

dianyang@sjtu.edu.cn, zhuyongxin@sjtu.edu.cn, ian_malcolm@sjtu.edu.cn, ylhe@sjtu.edu.cn, steph@nus.edu.sg

Abstract **— Water contaminants are among the critical sources of emerging pollutants, such as environmental endocrine concentrated in fishes and shrimps. Provenance information, describing the source and the route of transmission of pollutants, is important for environmental government to mitigate the threat of water contaminants in our urbanized society. However, existing provenance methods are not able to detect the source of these emerging pollutants due to movement of creatures who bio-magnifies concentrated environmental endocrine over a detectable level while the emerging pollutants in moving water are usually lower than a detectable level. To tackle the critical provenance issue, we propose an intelligent scheme by mining fish examination results or reservoir sensor readings along food chains to detect contaminants and provenance sources. Our scheme is capable of identifying the existence of contaminants, the possible sources as well as the provenance process.**

Keywords — data provenance; water contaminant; food chain; modelling

I. INTRODUCTION

As the urbanization process accelerates in most of the countries, overcrowded megacities are evolving into the dangerous condition of emerging pollutants among other prominent environmental issues. Water contaminants, have become important environmental pollution with increasing frequency of occurrences and broader damages. Some chemical products such as endocrine and heavy metal originate below the sensing level of monitoring detection, spread through biological enrichment and accumulate in higher organisms, finally threat to human health.

The existing method [1], such as deterministic method and probabilistic method are not able to detect the source of these emerging pollutants due to movement of creatures where bio-concentration takes effect. We endeavor to explore provenance techniques for an effective solution to trace emerging water contaminants.

Provenance tracing is supposed to be a promising technology to handle the threat of emerging pollutants such as water contaminants, atmospheric pollution control [2] and sound pollution [3]. Data Provenance is the information that describes the ancestry and history of uniquely identifying objects. Provenance provides a vision of connectivity for anything, at anytime and anywhere, which are necessary for tracing the pathway of pollutant [4][5].

Existing provenance techniques, however, also indicates limitations if they apply to water contaminant monitoring. In our previous work [6][7], we proposed an provenance model to tackle the problem of tracing the source of foodborne diseases in food supply chains (FSCs). By model real food supply chain, we analyze the spread of potential infected food in the market. However, the model in our previous is not sufficient to solve the problems in moving water contaminants because of the following reasons:

- In those models, it was assumed that the total number of food items is known, while the population of objects in nature is unknown.
- Besides the source of contaminants, we also need to know the total amount of pollutants.
- The concentrations of contaminants in the objects on the bottom of the food chain in nature, such as algae, are usually lower than a detectable level. Therefore the sampling method in our previous work is not suitable for the problem in this work.

In this paper, we establish a model of provenance in moving water food chain to resolve the limitations of provenance discussed above and simulate the water contaminant spread in Shanghai Qingcaosha Reservoir. This model of provenance with sensors provides good properties for government to trace the origin of contamination in water pollution. With the model, we propose a new provenance method to track contaminate source with a smart sensor data collection strategy. The contributions of this paper are as follows:

- The definition of a model of provenance in the moving water food chain as well as a data structure to organize related information about food chains.
- Design and implement of sampling algorithm to get a small and a representative portion of the samples from fishing operations. The sample rate is dynamic based on mark-recapture method.
- Design and implementation of a tracing algorithm to detect the origin of contamination by traversing the food chain from bottom to top. The contribution rate of species in the contiguous food chain is calculated

978-1-4799-8713-9/15 $31.00 © 2015 IEEE
DOI 10.1109/UKSim.2015.15

- through the Euclidean distance between sampling points.
- Simulation with the model under the constraints of the Shanghai Qingcaosha Reservoir to evaluate the model of provenance in several scenarios.

The rest of this paper is organized as follows. Section 2 presents related work. Section 3 proposed the model of provenance and corresponding algorithms. Section 4 presents the model simulation and analysis. Section 5 draws the conclusion.

II. RELATED WORK

In the context of water contaminants, information systems are vital to assist decision making in a short time frame, potentially allowing decisions to be made in real time. Deterministic method and probabilistic method are the major method tracing the source of water pollution[8].

Deterministic method refers to using mathematical physics equation to analyze pollutant movement, including direct analytical solution and simulation approaches. Direct solution, that is, by taking canonical transformations to convert inverse problems into a suitable problem for analytical and numerical solution.

Wei et al [9] designed a provenance model to solve spatial fractional anomalous diffusion equation based on regular optimum perturbation coupling method. Simulation optimization method is based on the difference between the measured value and the simulation value to get the optimal solution. Jha and Datta et al [10] implemented simulated annealing algorithm to underground water pollution provenance. Mo [11] established differential evolution algorithm model with single-point and multi-point stationary pollution source recognition.

Probability method solves the occurrence probability of a particular event. Cao et al [12] developed a mathematical model of convection-diffusion equation of pollution source use Bayes-Monte Carlo method. Chen et al [13] introduced water contaminant identification problem in Bayes-Monte Carlo approach. Cheng and Jia [14] studied river pollution provenance based on reverse probability.

However, deterministic method considers an error factor by applying perturbation analysis of results after calculating all pollution parameters. The probabilistic method avoids distortion of optimal parameters which brings risk in decision-making, though, contains strong randomness and the calculation will exponentially increase with the number of parameters grows [15][16][17][18].

In this paper, we proposed a numeric simulation model with which we combine food chain backtracking algorithm and mark-recapture sampling method in tracing contaminant sources and applied to a reservoir provenance issue.

III. MODELLING OF PROVENANCE

Due to the nature of the food chain, the provenance in a food chain is viewed as a directed acyclic graph (DAG), in which each node stands for one location keeps some batches of foods for a period.

A. Food chain provenance model

Generally, typical river food chain system includes: fish, shrimp, plankton and algae. To trace possible pollutant diffusion, environmental researchers usually capture creatures in similar categories at regular spatial and time intervals to analyze the pollutant content in the creatures' bodies. Sensor database stores sampling data and transmits to the data repository analyzing provenance through the food chain.

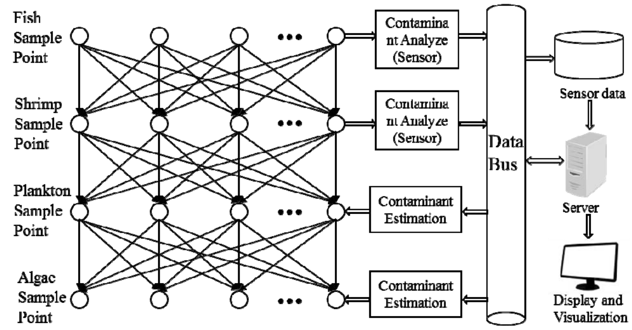

Fig. 1 An illustration of system's deployment structure for food chain provenance

B. Data structure

In the context of provenance, traceability refers to the capability to trace the entire path of the food chain: algae, plankton, shrimp and fish. Data structures of location and sampling information sets the foundation sufficient traceability.

Considering the need of the latter work, we defined two types of data structure in this section. All the two kinds of data will be stored in a central database.

Table I shows the information recorded for every sampling point location. Each point contains the distance of all other sampling points in the next layer of the food chain. The enrichment of contaminant has an intimate relationship with the distance of sampling point.

TABLE I. DATA STRUCTURE OF LOCATION INFORMATION

Distance	Fish1	Fish2	Fish3	…	Fishn
Sample Point1	Shrimp1	Shrimp1	Shrimp1	…	Shrimp1
Sample Point2	Shrimp2	Shrimp2	Shrimp2	…	Shrimp2
…	…	…	…	…	…
Sample Pointn	Shrimpn	Shrimpn	Shrimpn	…	Shrimpn

Table II records organism captured in each sampling point in the food chain during its procession.

TABLE II. DATA STRUCTURE OF SAMPLING POINT INFORMATION

Category	Fish	Shrimp	Plankton	Algae
Location	x	x	x	x
	y	y	y	y
Toxin	Fish_toxin	Shrimp_toxin	Plankton_toxin	Algae_toxin
Distance	Distance[n]	Distance[n]	Distance[n]	Distance[n]
Contribution rate	Contribution[n]	Contribution[n]	Contribution[n]	Contribution[n]

C. Analytical model of contribution rate

For the creature caught in each sampling point, its food source comes from its prey in the lower level of the food chain, locating in other sampling point. Thus, the contaminant in a creature's body composed of the food that creature preys on the reservoir. Assuming that the direction of movement of the predator is random, the proportion of prey exposes direct relevance to the distance of hunting.

For example, the fish in sampling point *Fish0* caught shrimp from all sample points *Shrimp0* to *Shrimp-n*. Due to the distance of each shrimp point is different, the food portion of fish in *Fish0* is various. The nearest shrimp point contributes most to fish point. Thus the contribution rate formula as follows:

$$Contribution_{ij} = \frac{\frac{1}{d_{ij}}}{\sum_{0}^{n} \frac{1}{d_{in}}} \tag{1}$$

Here, d is the Euclidean distance of a corresponding sampling point in the adjacent food chain. So the $Contribution_{ij}$ is the contribution rate of *Shrimp (j)* to *Fish (i)*. The scale of food source from one species to another can be calculated from Equation (1).

IV. PROVENANCE ALGORITHMS

The algorithms we proposed for the model of provenance consisted of two important parts: sampling from the whole collection of the underwater species; tracing the chain to detect the pollutant source.

A. Sampling algorithm

Since it is usually too expensive to test every individual species in the food chain, we apply mark-recapture sampling with the purpose to diminish the necessary number of sampling conducted before tracing and to retain the accuracy of our tracing algorithm. The pseudo-code of sampling algorithm is shown as Fig.2.

Here we first use mark-recapture sampling to the sum of fish in reservoir. Then the algorithm determines the appropriate sampling number. For example, in a reservoir with 5000 fish, 50 fish should be a suitable sample amount, so the percentage is 0.01. Later we catch fish in each sample point with fixed ratio. Peterson estimates the population quantity by the formula:

```
Input: K = Number of animals marked on the first visit,
        n = Number of animals captured on the second visit,
        k = Number of recaptured animals that were marked
Output: NS=Number of sampling fish
N=(n+1)*(K+1)/(k+1)-1
for percentage=0.01: #min_percentage do
    NS=N*percentage
    if (min_sample<NS<max_sample)
        break
    else
        percentage = percentage/2
end for
```

Fig. 2 Pseudo-code for sampling algorithm

$$\hat{N}_{Peterson} = (Kn)/k \tag{2}$$

B. Tracing algorithm

After sampling and testing, the tracing procedure checks the information stored for every sampling point and measures the contaminant content in each creature, including fish and shrimp. After that, the procedure sets the contaminant enrichment factor, according to the pollutant type and traces to the next layer of the food chain until the algae level. The algorithm, pseudo-code of the tracing procedure is shown as Fig.3.

```
Input: samples' spatial information and examination results
Output: contamination origin
for i=1:#food chain layer do
//In next iteration replace Fish by Shrimp, Plankton, and Algae
    for j=1:#shrimp sampling point do
        Detect contaminant content in fish body Fishtoxin_i
        for k=1:#shrimp sampling point do
            Distance_jk = sqrt((x_j - x_k)^2 + (y_j - y_k)^2)
            Distance_jk * = Distance_jk * t (adjust by water speed and direction)
        end for
        Contribution_ij =(1/Distance_jk)/( sum_0^{#sampling point} 1/Distance_jk )
    end for
    for j=1:fish sampling point do
        for k=1:shrimp sampling point do
            Shrimptoxin_k += Fishtoxin_j * Contribution_jk * α
        end for
    end for
end for
for i=1:#Algae sampling point do
    sort Algaetoxin_i from most to least
end for
origin = max(Algaetoxin_i)
```

Fig. 3 Pseudo-code for tracing algorithm

As a generic model, we suppose that the food follows a uniform distribution pattern in reservoir of natural condition, creatures can move in all directions to capture their prey. Meanwhile the water flow effect should be considered. For the same food located upstream and downstream with same distance, downstream food is easier to be captured. Thus water speed and direction have significant impact on the creature's hunting scope and the contribution rate of food. Origin Euclidean distance should take water effect into consideration. For example, given an x direction, the impacts are defined as:

$$\begin{aligned} Distance_{ij} &= t*(v+s) \\ Distance_{ij}^{'} &= Distance_{ij} - s*t \end{aligned} \tag{3}$$

Here v refers to the moving speed of creature in x direction, s is the speed of water flow. $Distance_{ij}^{'}$ is the normalized distance considered water flow factor, and t is the time we observed. Suppose the reservoir is a two-dimensional surface, in both x and y direction the water flow influence creatures' feeding process.

After getting normalized distance, tracing algorithm calculate pollutant in the next layer of the food chain. For example, the amount of pollutant content of a shrimp sampling point is calculated by the amount pollutant in fish point multiplied by contribution rate, then an enrichment factor α.

Although the contaminant concentration in algae keeps in a low-level, by biological enrichment affect, creatures in high-levels of a food chain accumulate large amount of toxic substance. Samples of creatures alone in low levels of the food chain, such as algae samples alone, are not able to determine water contaminant source. With our provenance process model, we can reveal the content data of water pollution and confirm the toxin source.

V. SIMULATION SETUP AND RESULTS

A. Dataset of simulation

In order to simulate the provenance model, we chose the data set of the Shanghai Qingcaosha Reservoir for a food chain network presented as Fig. 4.

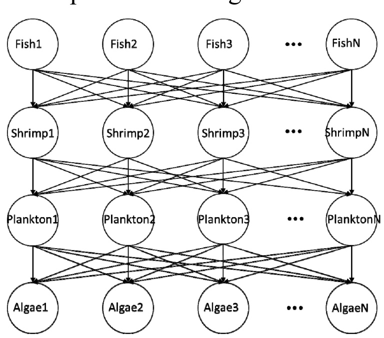

Fig. 4 The Topology Structure of food chain

In this case, we set 10 sampling points of each layer in different location of the Qingcaosha Reservoir to capture fish and shrimps. First level is fish sampling point from *Fish1* to *Fish10*. Second level is the shrimp sampling point from *Shrimp1* to *Shrimp10* and so on. For creatures like fish and shrimp, its moving features and water flow should be considered. But for algae, because it grows statically in the water, it does not consider the water flow effect.

Meanwhile, the content of toxic substance in the algae decides the source of pollutant, the closer algae located in pollutant source, the more toxic algae absorbs from the water. Thus the algae sampling point with most toxic content is deemed to be the source of pollutant.

B. Water pollution diffusion impact model

To simulate the impact of origin, the pollutant source is first set by the system. Suppose the reservoir is a two-dimensional surface, pollutant dispersion can be described by 2-D point-source diffusion simulation model [19]. The model of continuous point-source pollutant with water flow is shown as follows:

$$\begin{cases} D_x \dfrac{\partial^2 C}{\partial x^2} + D_y \dfrac{\partial C}{\partial y^2} - u_x \dfrac{\partial C}{\partial x} - u_y \dfrac{\partial C}{\partial y} - KC = 0 \ \ (\mathrm{x} > 0, \mathrm{y} > 0) \\ Boundary\, condition: \left.\dfrac{\partial C}{\partial y}\right|_{y=0} = 0 \end{cases} \tag{4}$$

The analytical solution of the model is:

$$C(x,y) = \frac{m}{4\pi(\frac{x}{u_x})^2 \sqrt{D_x D_y}} \exp[-\frac{(y-u_x y/u_x)^2}{4D_y x/u_x}] \exp(-\frac{Kx}{u_x}) \tag{5}$$

For river or reservoir that flows gently with contiguous depth, the parameters u_y, D_x can be dismissed. Thus the Equation (5) is equivalent to the following expression:

$$C(x,y) = \frac{m}{2u_x H \sqrt{\pi D_y x / u_x}} \exp[-\frac{(u_x y)^2}{4D_y x}] \exp(-\frac{Kx}{u_x}) \tag{6}$$

Here, m is the weight of total pollution in unit time, and H refer to average water depth.

As a simulation example, if there is a release of 200KG pollutants at the source of reservoir (near sampling point1) and after 2 months, the pollutant expanded in the reservoir, got absorbed by algae and made impacts on the food chain. The concentration of pollutant is distributed as Fig. 5.

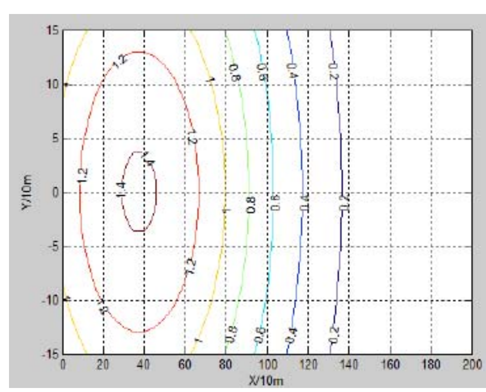

Fig. 5 The concentration distribution of continuous point-source pollutant

C. Provenance tracing results

For tracing algorithm part, the system is simulated and tested under the pollutant provenance of the food chain. The results of the algorithm are listed in Fig. 6 and Fig. 7.

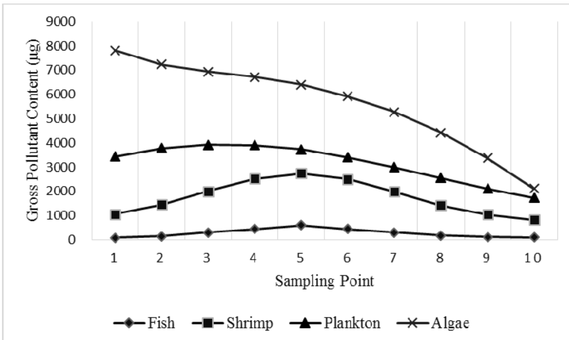

Fig. 6 Gross pollutant content in each layer of food chain

Fig. 6 reveals the distribution of pollution amounts in each sampling point. The pollutant content accumulates through the food chain from the top layer (fish layer) to bottom (algae layer). The algae layer possesses substantial contaminant with relatively low concentration.

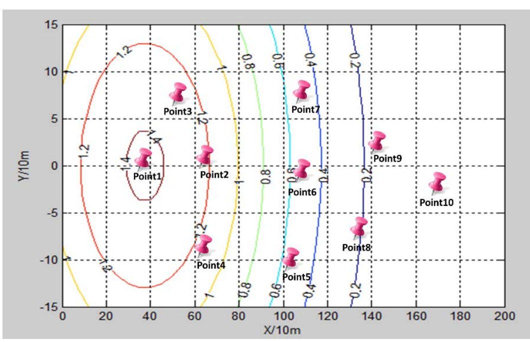

Fig. 7 The sampling point location in contaminated water

In the meanwhile, the content decreases from *Algae Point1* to *Algae Point10*. In other layers, the maximum value of gross pollutant content moves forwards, indicates that the toxic passes through predator-prey relations from upstream to downstream. Therefore, bioaccumulation of pollutant content corresponds to water pollution diffusion, which is shown in Fig. 7.

Table III shows the concentration of pollutant in each layer. Comparing Table III and Fig. 5, we can observe that even though algae layer preserves most of pollutants in terms of gross pollutant content, its concentration within each sample of algae keeps in a very low level. It explains that chemical detection experiment with only samples in lower layers of the food chain is not able to detect water pollution. Bio-concentration effect concentrates contaminants in high level creatures, which are detectable with samples in the upper layers of food supply chain.

TABLE III. POLLUTANT CONCENTRATION IN EACH LAYER OF FOOD CHAIN

Samp-ling Point	Fish (ppm)	Shrimp (ppm)	Plankton (ppm)	Algae (ppm)	Water (ppm)
1	2.1468207	0.4246792	0.0442724	0.0011339	0. 000 0030
2	1.5446586	0.4219522	0.0307245	0.0019256	
3	2.0528562	0.4924437	0.0495427	0.0021899	
4	2.1292028	0.5872378	0.0460105	0.0025534	
5	2.0869802	0.5961036	0.0319130	0.0021502	
6	2.3486036	0.4431503	0.0485268	0.0021243	
7	1.8343030	0.5262405	0.0374822	0.0026640	
8	2.4236364	0.5534076	0.0463184	0.0025925	
9	1.7531034	0.5593026	0.0378677	0.0021162	
10	1.6283161	0.5526089	0.0456779	0.0028726	

With detectable samples in the upper layers of food supply chain, result of tracing simulation demonstrates that by tracing the food chain from the top to bottom, toxic substance was found to concentrate at the *Sampling Point1*. It means most of contaminant diffusion originated from algae around *Sampling Point1* which is the source of pollution.

D. Additional impacts of water flow and sampling intervals

In the reservoir, the speed and direction of water affect the movement of a creature. For food with same distance, downstream food is easier to be obtained than upstream food. Assume two shoal of shrimp keep one kilometer from the fish sampling point, one in downstream and another on upstream. The relation of food contribution of two shoals and water speed is shown in Fig. 8. With the water speed increase, contribution rate of downstream increase while the upstream decrease. In extreme cases, the upstream contribution rate would fall to zero if water speed equals to fish swimming speed. These changes affect the simulation model of concentration by replacing the analytical expression in Equation (1) with numeric values of concentration rates.

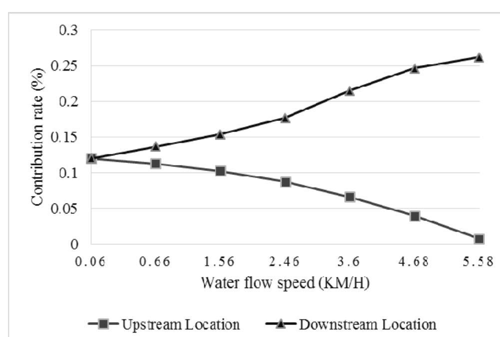

Fig. 8 The impact of water speed on contribution rate

The area of Qingcaosha Reservoir covers about 70 square kilometers, which makes it infeasible to deploy sensors or sampling in high density. To have an initial evaluation of the impact of distance between sampling points, we equally set 10 sampling points of each layer to measure the provenance. The distance of food location is associated with contribution rate. For a fish sampling point, the contribution rate of a shrimp sampling point varies with distance from 0.5km to 70km. The result is shown in Fig. 9, indicating that sampling interval needs to be less than 5KM (50*0.1KM) to discover the effective concentration rate.

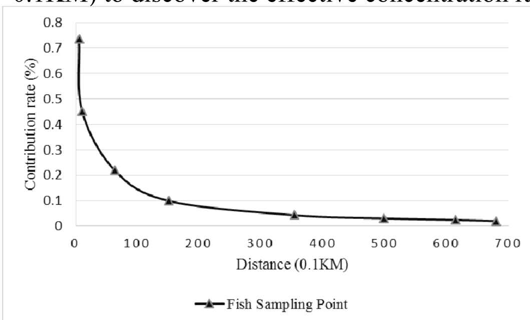

Fig. 9 The impact of distance between sampling point on contribution rate

VI. CONCLUSION

In this paper, we proposed a provenance model and simulate emerging water contaminants to handle the properties of moving creatures and contaminant bio-magnification. We designed and implemented a tracing algorithm on top of the model to find the source of contamination through water food chain. With consideration of water pollution diffusion, we simulated the process of pollution with constrained parameters from the Shanghai Qingcaosha Reservoir. Through the contribution rates, we analyzed the water flow speed and direction factor that would affect the spread of toxic substance. It was observed in simulation results that our provenance model and algorithms are able to trace the source of contamination.

Our future work is to further make practical implementation of the provenance of food chain based on cloud storage services as we assumed that all provenance information of creatures was hosted in a centralized database and these provenance meta-data are organized in a uniform manner in this paper.

ACKNOWLEDGEMENT

This paper is sponsored in part by the Shanghai International Science and Technology Collaboration Program under Grant 13430710400, National Natural Science Foundation of China (61373032), and National Research Foundation (NRF), Prime Minister’s Office, Singapore under its Campus for Research Excellence and Technological Enterprise (CREATE) program by Shanghai Jiao Tong University (SJTU) and National University of Singapore (NUS).

REFERENCES

[1] Zeng, L., Shi, L., Zhang, D., & Wu, L. A sparse grid based Bayesian method for contaminant source identification. Advances in Water Resources, 37, 1-9. (2012)

[2] Shelke R, Kulkarni G, Sutar R. Energy Management in Wireless Sensor Network. Computer Modelling and Simulation (UKSim), 2013 UKSim 15th International Conference on. IEEE, 2013: 668-671.

[3] Aziz A, Hossain M. Inherent inter-vehicle signaling using radio frequency and infra-red communication. Computer Modelling and Simulation (UKSim), 2012 UKSim 14th International Conference on. IEEE, 2012: 211-215.

[4] Bauer, S., & Schreckling, D. Data Provenance in the Internet of Things. EU Project COMPOSE, Conference Seminar(2013)

[5] Pang, Z., Chen, Q., Han, W., & Zheng, L. Value-centric design of the internet-of-things solution for food supply chain: value creation, sensor portfolio and information fusion. Information Systems Frontiers, 1-31. (2012)

[6] Zhang, Q., Wang, D., Huang, T., Zhu, Y., Qiu, M., Ni, M., & Xie, G. Modelling provenance in food supply chain to track and trace foodborne disease. In International Conference on Computer Modeling and Simulation (pp. 69-75). (2012)

[7] Zhang, Q., Huang, T., Zhu, Y., & Qiu, M. A Case Study of Sensor Data Collection and Analysis in Smart City: Provenance in Smart Food Supply Chain. International Journal of Distributed Sensor Networks. (2013)

[8] Yang, H., Xiao, Y., Wang, Z., Shao, D., & Liu, B. On source identification method for sudden water pollution accidents. Advances in Water Science. (2014)

[9] Wei, H., Chen, W., Sun, H., & Li, X. A coupled method for inverse source problem of spatial fractional anomalous diffusion equations. Inverse Problems in Science and Engineering; Formerly Inverse Problems in Engineering, 18(7), 945-956. (2010).

[10] Jha, M., & Datta, B. Three-dimensional groundwater contamination source identification using adaptive simulated annealing. Journal of Hydrologic Engineering, 18(3), 307-317. (2012).

[11] Mou X., Research on inverse problem of pollution source term identification based on differential evolution algorithm. Chinese Journal of Hydrodynamics, 1, 005. (2011).

[12] Cao, X., Song, J., Zhang, W., & Zhang, L.. MCMC method on an inverse problem of source term identification for convection-diffusion equation. Chinese Journal of Hydrodynamics, 2, 003. (2010).

[13] Chen, H., Teng, Y., Wang, J., Song, L., & Zhou Z., Event Source Identification of Water Pollution Based on Bayesian-MCMC. Journal of Hunan University (Natural Sciences), 6, 015. (2012).

[14] Cheng, W. P., & Jia, Y. Identification of contaminant point source in surface waters based on backward location probability density function method. Advances in Water Resources, 33(4), 397-410. (2010).

[15] Amirov, A., & Ustaoglu, Z.. On the approximation methods for the solution of a coefficient inverse problem for a transport-like equation. Computer Modeling in Engineering and Sciences (CMES), 54(3), 283. (2009)

[16] Chen, W., Cheng, J., Lin, J., & Wang, L.. A level set method to reconstruct the discontinuity of the conductivity in EIT. Science in China Series A: Mathematics, 52(1), 29-44. (2009)

[17] Liu, C. S., Chang, C. W., & Chang, J. R. A new shooting method for solving boundary layer equations in fluid mechanics. Computer Modeling in Engineering & Sciences, 32(1), 1-15. (2008).

[18] Liu, X., Yao, Q., Xue, H., Chu, K., & Hu, J. Advance in inverse problems of environmental hydraulics. Advances in Water Science, 20(6), 885-890. (2009).

[19] Song X. S. The application of MATLAB in environmental sciences. Chemical Industry Press. 2008