Using robust Bayesian network to estimate the residuals of fluoroquinolone antibiotic in soil

Xuwen Li¹ · Yunfeng Xie² · Lianfa Li³ · Xunfeng Yang³ · Ning Wang¹ · Jinfeng Wang³

Received: 16 January 2015 / Accepted: 18 May 2015 / Published online: 5 July 2015
© Springer-Verlag Berlin Heidelberg 2015

Abstract Prediction of antibiotic pollution and its consequences is difficult, due to the uncertainties and complexities associated with multiple related factors. This article employed domain knowledge and spatial data to construct a Bayesian network (BN) model to assess fluoroquinolone antibiotic (FQs) pollution in the soil of an intensive vegetable cultivation area. The results show: (1) The relationships between FQs pollution and contributory factors: Three factors (cultivation methods, crop rotations, and chicken manure types) were consistently identified as predictors in the topological structures of three FQs, indicating their importance in FQs pollution; deduced with domain knowledge, the cultivation methods are determined by the crop rotations, which require different nutrients (derived from the manure) according to different plant biomass. (2) The

performance of BN model: The integrative robust Bayesian network model achieved the highest detection probability (pd) of high-risk and receiver operating characteristic (ROC) area, since it incorporates domain knowledge and model uncertainty. Our encouraging findings have implications for the use of BN as a robust approach to assessment of FQs pollution and for informing decisions on appropriate remedial measures.

Keywords Antibiotic residue · Bayesian network · Fluoroquinolone antibiotic · Intensive vegetable cultivation area

Introduction

Veterinary antibiotics are probably the most successful family of drugs developed so far to treat and prevent disease and protect animal health worldwide. They are also incorporated into animal feed to improve growth rate and feed efficiency. The significant increase in the number of large-scale animal feeding operations for swine, poultry, and cattle has resulted in a considerable amount of antibiotics commonly added to animal feeds (Van Doorslaer et al. 2014). Antibiotics are poorly adsorbed in animal guts, resulting in about 30–90 % of the parent compound excreted unchanged in feces and urine (Sukul and Spiteller 2007). Some antibiotics remain biologically active after entering in the environment.

In organic vegetable production, the application of chemical fertilizers is limited throughout the intensive vegetable cultivation fields. Consequently, the vast majority of manure with certain antibiotic residues is applied to agricultural land as a nutrient source for plants (Li et al. 2013; Zhao et al. 2010). After entering the soil, antibiotics can be transported to surface and ground water and taken up by plants (Kuchta and Cessna 2009). There is evidence of ecotoxic antibiotic effects on plants

Responsible editor: Marcus Schulz

Lianfa Li and Jinfeng Wang contributed equally to this work.

Electronic supplementary material The online version of this article (doi:10.1007/s11356-015-4751-9) contains supplementary material, which is available to authorized users.

✉ Lianfa Li
lsatial@gmail.com

✉ Jinfeng Wang
wangjf@lreis.ac.cn

¹ Department of Environment and Health, School of Public Health, Shandong University, Jinan 250012, China

² State Key Laboratory of Environmental Criteria and Risk Assessment, Chinese Research Academy of Environmental Sciences, Beijing 100012, China

³ State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, A11, Datun Rd, Anwai, Beijing 100101, China

(Aristilde et al. 2010; Jin et al. 2009) and soil microorganisms (Girardi et al. 2011). Furthermore, it is anticipated that directly and potentially adverse impacts like as allergic/toxic reactions, chronic toxic effects due to prolonged low-level exposure, development and spread of antibiotic-resistant bacteria, and disruption of digestive system functioning will occur to human (Kumar et al. 2005; Martinez 2008; Xiao et al. 2008). Therefore, antibiotics contamination in a vegetable planting area has become an urgent problem (Smukler et al. 2008).

The manure application is the sole source inputs, and antibiotic residue levels in soil of greenhouses were presumed to be closely related to the manure usage and species. However, our previous research emphasized the crop rotations is the major determinant of spatial stratification of the fluoroquinolone antibiotic (FQs) in soil (Li et al. 2013). This result indicates some complicated relationships existing in these influencing factors. Due to food safety issues, because the vegetable can absorb the FQs antibiotics from the soil and transport them into plant tissues (Hu et al. 2010), it is necessary to clarify the causes of antibiotics residues in soil. Investigating the environmental and landuse conditions that lead to antibiotic soil pollution is critical to develop potential approaches to prevent soil pollution.

In addition, the transformations among FQs make their distributions in soil more complicated. Like as the substance of enrofloxacin (EFX), is a pyridone carboxylic acid derivative, could be de-ethylated to its primary metabolite, ciprofloxacin (CFX), which also displays potent antimicrobial activity (Garcia et al. 2005). Therefore, choosing a suitable method to demonstrate their relationship is of the utmost importance. Statistical methods are commonly applied to quantify causal relationships between drivers (e.g., environmental and landuse variables) and outcomes (e.g., pollution). Bayesian network (BN) is one particularly attractive approach to relating the environmental and landuse conditions to antibiotics contamination of soil. BN enables us to combine any relevant statistical/quantitative data available with qualitative data and subjective judgments (Li et al. 2010, 2012). With BN, it is possible to quantify expert beliefs about the dependencies among different variables and to propagate consistently the impact of evidence on the probabilities of uncertain outcomes. Therefore, in this study, we aim to estimate the residue level of FQs in soil and investigate the relationship between soil pollution levels and planting patterns using robust Bayesian Network and other probabilistic models.

Materials and methods

Study area

The study area is located at southeast of Shouguang county, Shandong province of China, lying between north latitude

36°45' N and 36°52'8.4" N, and between east longitude 118°48'54" E and 118°59'49.2" E. Shouguang county is the largest vegetable producing county in China, with intensive, regional, and consistent crop rotations. A large quantity of manure is used in this area, which has caused antibiotics residues in soil-vegetable system (Li et al. 2014).

The study selected 100 vegetable greenhouses with a typical spacing of about 1 km between individual greenhouses. Using a small shovel, soil samples were collected 0–15 cm below the soil surface during November 2010. The soil samples were all collected on soil ridges that were about 15–20 cm higher than the surface soil and contained almost no roots. This was done in order to avoid the effect of plant uptake in later analysis (it is difficult to calculate levels of antibiotics in the roots, leaves, fruits, and other biomass of different vegetables in field research). Samples were collected at five sites chosen within each greenhouse following an S-shape and were combined and fully homogenized to form one sample. These samples were immediately transported to the laboratory under cooled conditions for analysis. The soil type in the target area is mainly cinnamon soil (Alfisol). The physical and chemical properties of one sample are described as follows: soil organic matter was about 32.7 g/kg; pH, 6.7; CEC, 17.4 cmol/kg; and the texture was approximately 45.2 % for clay (<0.001 mm), 34.6 % for silt (0.001–0.01 mm), and 20.2 % for sand (>0.01 mm).

FQs analysis

The concentrations of three kinds of fluoroquinolones including ciprofloxacin (CFX), enrofloxacin (EFX), and norfloxacin (NFX) in soil were analyzed using high-performance liquid chromatography and fluorescence detection (HPLC–FD) with the method developed by Turiel et al. (2006). Lyophilized soil (1.0 g) was placed in a 10 ml centrifuge tube, and the FQs were extracted using 5 ml of potassium phosphate buffer and acetonitrile ($v/v=1:1$). The centrifuge tubes were vortexed for 10 min, sonicated for 15 min, and then centrifuged at 3000×g under air-cooled conditions. The extraction procedure was performed 3 times, and the resultant supernatants were combined.

Oasis HLB (200 mg, 6 ml) and SAX cartridge (200 mg, 3 ml) were obtained from Waters Corporation (Milford, MA, USA) and Bonna–Agela Technologies (Tianjin, China). Each cartridge was sequentially pre-conditioned with 6.0 ml methanol and 6.0 ml Milli-Q water. The mixture extraction solution with organic reagent under 5 % was loaded onto the cartridge. The cartridge was then rinsed with 6 ml of Milli-Q and lyophilized for 40 min. Elution of the antibiotic compounds from the cartridges was done with 6 ml of methanol. Methanol eluate was evaporated to near dryness under a gentle stream of nitrogen, followed by redissolution with 1.0 ml of methanol. The mobile

phase consisted of acetonitrile and 0.01 mol/L oxalic acid (80:20 v/v) at 1 ml/min flow rate.

Their respective recoveries of CFX, EFX, and NFX were about 71–82 %, 68–85 %, and 65–87 % in soil samples. External calibration curves were constructed by preparing the standard solutions at six known concentrations (0.01–1.0 mg/kg for FQs). Concentrations of the analyses in sample matrices were determined by using peak areas which corresponded to the unknown concentrations in the calibration curve. The detection limits of the three antibiotics were lower than 0.1 µg/kg, and the relative standard deviation (RSD) was in the range of 2.5–5.0 %. Noticeably, fluoroquinolones (FQs) were detected in all soil samples.

Data set

Target variables

We constructed a BN model for CFX, NFX, and EFX, with their concentration in soil as the target variables. As the concentration of CFX and NFX is much higher than that of EFX (Table S1), we chose different reference value of “high risk” for CFX, NFX, and EFX. The target variables were classified as low risk or high risk (to be modeled in BN since BN just takes the qualitative variable as the target one) using 50 µg/kg as the threshold value for CFX and NFX and 5 µg/kg for EFX. As a result, 58 greenhouses were classified as high risk by CFX, 37 greenhouses were classified as high risk by NFX, and 57 greenhouses were classified as high risk by EFX. We also used the other different thresholds such as mean and median to divide the low and high degree of concentrations and conduct the sensitivity analysis. The analysis showed that our threshold values mentioned above achieved the best classification performance although the differences between different division schemes were small.

Factors associated with FQs pollution

The environmental and landuse conditions (formally considered as statistical risk factors potentially affecting FQs pollution in vegetable greenhouses) include cultivation methods, crop rotations, manure types, manure quantity, greenhouse area, and topographic elevation. Cultivation methods, crop rotation, manure types are categorical variables, while the other three factors are continuous variables. We distinguish five crop rotations: “tomato–tomato,” “pepper–melon,” “pepper–eggplant,” “cucumber–cucumber,” and “melon–leaf vegetables”; three manure types: fresh manure (unprocessed), dry manure (fresh manure that has been fermented and dried), and mixed manure (fresh and dry); and two cultivation methods: plastic arched shed and solar greenhouse. The detail information has been demonstrated in our former research (Li et al. 2013).

Bayesian network

Bayesian network (BN) is a model of probabilistic inference that involves different nodes of random variables and their connections of probabilistic dependence. Thus, we can define a Bayesian Network B over the set of variables, V as an ordered pair (B_S, B_P) . The network B consists of two parts, the network structure, $B_S = G(V, E)$, and the assessment functions, $B_P = \{\gamma_u : \Omega_u \times \Omega_{\pi_u} \rightarrow [0 \dots 1] | u \in V\}$. The network structure is a directed acyclic graph ($E \subseteq V \times V$ is the set of directed edges, representing the probabilistically conditional dependence relationship between random variable. Such a probabilistic dependency relationship may represent a cause-effect or diagnostic relationship (Russell and Norvig 2010). A diagnostic relationship leads from observed effects to possible or hidden cause. In the set of assessment functions, Ω_u is the state space, the finite set of values for the variable u , π_u is the set of parent nodes of variables for u , indicated by B_S ; Ω_X is the Cartesian product of all state spaces of the variables in X ; γ_u uniquely defines the joint probability distribution $p(u|\pi_u)$ of u conditional on its parent set, π_u (Li et al. 2010).

Construction of the structure and solution of the network parameters (conditional probabilistic tables) are the major steps for construction of BN. We can build the network structure by data learning and domain knowledge. To learn the network from the data, we can apply different heuristic (K2, hill climbing, and Tree Augmented Naive) or general-purpose (Tabu and simulated annealing) search strategies (Table 1).

A self Bayesian network is a BN based on data learning and refined using domain knowledge about the known dependent relationships between random variables. For example, in Fig. 1, the basic network was learned using “Tan” algorithm from data but we added extra connections from cultivation method to crop rotation and from crop rotation to chicken manure type according to domain knowledge introduced earlier.

Once the network model is setup, we can use the probabilistic inference to make the prediction for the target variable. For example (Fig. 1a), the “NFX” node is the random target variable and its state space Ω_c be binary, then $\Omega_c = \{\text{“low”}, \text{“high”}\}$. In a specified BN, if some evidences are given, we can get the posterior probability or belief [abbreviated as *Bel* in Eq. (1)] of $t = \text{“NFX”}$ as the low or high level by calculating the marginal probability or using approximate inference:

$$\text{Bel}(\text{level} = \text{“low” or “high”}) = \sum_{u_i \in V, u_i \neq t} p(u_1, u_2, \dots, u_n; t) \quad (1)$$

where $p(u_1, \dots, u_n) = \prod_{u_i \in V} p(u_i | \pi_{u_i})$ is the joint probability over V .

Table 1 Learning algorithms for the network structure

Method	Description
K2	K2 (Cooper and Herskovits 1992) adds arcs with a fixed topological ordering of variables. In this method, the ordering is initially set as a naive Bayes network in which the target class variable is made the first in the ordering.
Hill climbing (HC)	Hill climbing (Buntine 1996) adds and deletes arcs with no fixed ordering of variables. This procedure is iterated until the highest value of the measure score (e.g., local Bayes) is obtained.
Tabu	Tabu search (Bouckaert 1995) is an optimal hill climbing algorithm. This algorithm applies a Markov Blanket correction to the network structure after a network structure is learned.
Tree augmented naive (TAN) Bayes	In Tree Augmented Naive Bayes (Friedman and Geiger 1997), the tree is formed by calculating the maximum-weight spanning tree using Chow and Liu's algorithm (Chow and Liu 1968).
Simulated annealing (SA)	Simulated annealing (Bouckaert 1995) randomly generates a candidate network close to the current network B_S . It accepts the network if it is better than the current one.

Robust BN

To mitigate the sampling bias and model uncertainties (also avoiding the overfitting problem), we use Bayesian model averaging (BMA) and Occam's Window (Morales-Casique et al. 2010) to produce a robust prediction of the levels of FQs pollution.

While BMA can average the predictions of the learning algorithms to reduce uncertainty, we can also use Occam's Window to select the qualified models and remove those poorly qualified models, thus improving the computation efficiency. Occam's Window (Hoeting et al. 1999) has two principles: (1) if a model receives much less support (e.g., the ratio of 20:1) than another model with maximum posterior probability, then it should be dropped; (2) complex models that receive less support than their simpler counterparts should be dropped.

We use six methods described above for construction of Bayesian network, including five data learning search algorithms (K2, hill climbing, Tan, Tabu, simulation annealing; Table 1) and oneself method that combined data leaning and domain learning to get the local structures of Bayesian network. Then, the BMA and Occam's Window were used to select the highly qualified models and average them to get the robust predictions that can decrease the models' bias and uncertainty. This systematic method integrated over multiple single BNs is called robust BN in our study.

Optimal discretization and feature selection

Optimal discretization

Discretization is employed to transfer the continuous variables to discrete variables for use in BN since BN contains only qualitative or discrete variables for modeling and inference. We obtain the discrete intervals according to domain knowledge or use a discretization algorithm to discretize continuous factors such as manure quantity, greenhouse

area, and topographic elevation. By discretization, we can use quantitative factors along with qualitative or discrete factors to strengthen the model's predictability of the residual level of FQs antibiotics in BN. If little domain knowledge for discrete intervals is known, we can use discretization algorithm that can be used to make an automatic division. The algorithm we used in this research is designed according to the binary "recursion" idea and the principle of optimization (the minimal description length (MDL) stopping criteria) in Fayyad and Irani's algorithm. This discretization method was reported in our previous article (Li et al. 2010). Readers can refer to the article for technical details.

The optimal splits for discretization of continuous variables were calculated, and the results were shown in Table 2. For NFX, there was no split due to insignificance of numeric features. While for EFX, the elevation was discretized into two parts (below and above 11.5 m) and other variables were not discretized due to their insignificance. For CFX, elevation (three intervals: [8 m, 14 m], [14 m, 24 m], [24 m, 47 m]) and use volume for each greenhouse (below and above 1750 kg) were discretized by the threshold values.

Feature selection

For feature selection, we use the information-based feature selector (Quinlan 1996): Information Gain Ratio (GR). GR takes into account the information that each feature contains and measures the gain ratio given the feature to be assessed.

We used information gain ratio (GR) to measure the contribution of the feature to the classification of the target variable (low or high residuals of antibiotics). GR ranges from 0 to 1 with higher value indicating higher contribution and priority of choice of the feature to the classification. Table 3 shows GR for each feature (each predictive variable) in the models, and generally, the feature with $GR > 0$ can be selected into the models and those with GR close to 0 can be filtered as useless features.

Fig. 1 Bayesian network topology of NFX (a), EFX (b), and CFX (c) pollution

Table 2 Optimal splits for discretization for numeric features

Target variable	Optimal splits for discretization
NFX	No splits for insignificance of numeric features
EFX	Elevation: ≤ 11.5 , > 11.5 ; other not discretized due to insignificance
CFX	Elevation: [8, 14], [14, 24], [24, 47]; use volume for each greenhouse: ≤ 1750.0 , > 1750.0

Table 3 Optimal features selected

Target variable	Features selected
NFX	No selection of numeric features due to their insignificance. Selected features: chicken mature type 0.133; vegetable type 0.046; greenhouse type 0.002
EFX	Chicken mature type 0.016; vegetable type 0.058; greenhouse type 0.018; elevation: 0.178
CFX	Chicken mature type 0.067; vegetable type 0.112; greenhouse type 0.083; use volume for each greenhouse 0.194; elevation 0.181

Evaluation and prediction

To evaluate our method, we compared it with some other methods using scalar performance measures.

Compared methods

Our methods (self Bayesian network and robust Bayesian network) were compared with other prediction methods, including five BNs based on five search algorithms (Table 3) and seven other commonly used methods (Table 4).

Evaluation measures

We use two scalar measures, i.e., pd and ROC area.

- pd refers to the detection probability of high risk: It measures the proportion of correctly predicted positive instances among the actually positive ones. If a method achieves a higher pd, it can detect more positive instances (e.g., more cell units of high level of FQs pollution detected).
- ROC area is the area between the horizontal axis and the receiver operating characteristic (ROC) curve; it gives a comprehensive scalar value representing the model's expected performance. The ROC area is between 0.5 and 1, where a value close to 0.5 is less precise, while a value close to 1.0 is more precise.

Prediction Map by kriging and Thiessen–Voronoi diagram

Finally, we use the modeled robust Bayesian network to make a prediction of the high-risk probabilities of FQs pollution of the 100 selected vegetable greenhouses. Furthermore, kriging and Thiessen–Voronoi diagram are adopted to interpolate the surface of risk probabilities of FQs pollution for the whole study area based on the Bayesian predictions. Using kriging method, we can see spatial distribution and the trend of high-risk probabilities of FQs pollution. The Thiessen–Voronoi diagram is chosen to avoid overfitting problem by kriging, but it cannot present the trend. Kriging interpolation was carried out using R 3.0.1 software, and ArcGIS 10.1 was used to create a Thiessen–Voronoi diagram.

Results

Bayesian networks by data learning and domain knowledge

In Fig. 1, the arrow illustrates a probabilistic, causal, or influential link between two nodes. Each of the nodes (single-line or double-line ellipse) is a random variable (used in BN to represent an influential factor or target variable), and the direction of the connections (including the single or double straight lines) represents the probabilistic dependency (cause-effect or diagnostic)

Table 4 Prediction methods compared

Method	Description
J48	A C4.5 (Quinlan 1996) decision tree recursively partitions the training data by means of attribute splits and generates a pruned or unpruned tree using the information-theoretical concept of entropy.
Logistic regression (LR)	Logistic regression is a technique of probability estimation based on maximum likelihood estimation (Hosmer and Lemeshow 2000).
Normalized Gaussian radial basis function network (RBF)	Normalized Gaussian radial basis function network (Bugmann 1998) uses the k-means clustering algorithm to provide the basis functions and learns either a logistic regression (discrete class problems) or linear regression (numeric class problems) on top of that.
SMO	SMO (Keerthi et al. 2001) is an algorithm of support vector machine classifier that globally transforms nominal attributes into binary ones and multiclass problems are solved using pairwise classification.
Naive Bayes (NB)	Naive Bayes (John and Langley 1995) assumes that the predictive factors are conditionally independent given the target/dependent variable.
Random forest	A forest of random trees (Breiman 2001) is a meta-learning schema that embodies several base classifiers (CART) that are built independently and participate in a voting procedure to obtain a final class prediction.

relationships between paired variables. For example, the link with the arrow from FQs to crop rotations and cultivation methods represents a diagnostic link: according to the FQs, we can make a probabilistic inference for the status of crop rotations and cultivation methods. Firstly, according to the Bayesian structures of three antibiotics, the common factors of cultivation method, crop rotation, and chicken manure type were selected as predictors of FQs residues in soil. Secondly, although the three selected FQs belong to the same category and derive from the same manure resource, their selected predictors differ. The BN topology of NFX focuses on the common factors mentioned above, whereas in the structures of EFX and CFX, elevation is selected as a predictor, and the volume of antibiotic applied to each greenhouse is only selected in the BN of CFX. CFX has the highest concentration level and the most complicated relationship of all the predictors. As expected, the analysis showed that the quantitative factors such as manure quantity and greenhouse area played unimportant roles in FQs residues. The common factors such as cultivation method, crop rotation, and chicken manure type are consistently identified in their topology structures, indicating their pivotal status.

Comparative performance of multiple learners

Table 5 presents the ROC area of each method for prediction of FQs pollution levels in soil. The results show that the robust BN gave the best predictive performance, with much higher ROC area than other methods (from 5 % to more than 40 % improvement over other models). The performance of the self BN (combination of data learning and domain knowledge) is just secondary to the robust BN. The results show that consideration of domain knowledge and uncertainty is very important for improved model performance.

While robust BN and self BN showed the greatest improvements in total predictive performance, they still achieved the best predictions of high levels of FQs pollution (one major

concern associated with FQs pollution), as shown in Table 6. The results also demonstrate the importance of using domain knowledge to refine the network structure in BN and the consideration of uncertainty for classification, prediction, and identification of high levels of FQs pollution.

Finally, we used a kriging predictor and Thiessen–Voronoi diagram to assess the spatial distributions of these probabilities for the whole study area. Figures 2 and 3 show maps of the high-risk probabilities of NFX, CFX, and EFX pollution by kriging predictor and Thiessen–Voronoi diagram, respectively.

Discussion

FQs pollution and contributory factors

FQs residues are present in soils used for intensive cultivation of vegetables. Although CFX, EFX, and NFX belong to the same category and have the same skeleton structure, they are present at different concentrations (CFX>NFX>EFX).

The reasons for the differing concentrations of FQs residues in soil are complicated. Besides the differences between FQs species and concentrations in manure, this phenomenon might be related to the transformation of FQs during the entire process of chicken breeding and metabolism, manure storage, and their biochemical reactions in soil. Previous studies have shown that FQs could be converted into their epimers in different environments. For example, transformation from EFX to CFX was reported in plant root systems (Liu et al. 2013); the biotransformation of CFX and NFX by microorganisms was confirmed (Parshikov et al. 2001; Sukul and Spiteller 2007), and the same transformation was even reported in animals (Bidgood and Papich 2005; Fu et al. 2008). The transformation to CFX altered the original distributions of EFX and NFX in manure and soil matrix. Additionally, the transformation to

Table 5 Performance comparison of ROC areas

Model	NFX	CFX	EFX
J48	0.69	0.65	0.51
Logistic	0.70	0.80	0.52
RBF	0.69	0.79	0.56
SMO	0.64	0.77	0.69
Naive Bayes	0.70	0.80	0.54
Random forest	0.72	0.72	0.51
BN-K2	0.72	0.79	0.59
BN_Hill	0.70	0.76	0.57
BN-Tabu	0.65	0.71	0.58
BN-Tan	0.70	0.79	0.55
BN-SA	0.68	0.82	0.53
BN-self	0.80	0.87	0.69
Robust BN	0.81	0.90	0.74

Table 6 Performance comparison probability of detection for high concentration: pd

Model	NFX	CFX	EFX
J48	0.43	0.84	0.72
Logistic	0.48	0.83	0.56
RBF	0.41	0.84	0.58
SMO	0.49	0.84	0.61
Naive Bayes	0.45	0.83	0.60
Random forest	0.43	0.79	0.54
BN-K2	0.43	0.81	0.56
BN_Hill	0.38	0.78	0.35
BN-tabu	0.49	0.70	0.54
BN-Tan	0.49	0.83	0.63
BN-SA	0.48	0.84	0.79
BN-self	0.51	0.86	0.88
Robust BN	0.59	0.91	0.95

A

A

B

B

C

C

Fig. 2 Robust prediction maps of high-risk probabilities of NFX (a), EFX (b), and CFX (c) using kriging predictor

Fig. 3 Robust prediction maps of high-risk probabilities of NFX (a), EFX (b), and CFX (c) using Thiessen-Voronoi diagram

CFX is also one of the reasons that CFX is present at the highest concentration in soil (Fig. 1) and has a more complicated topology structure than EFX and NFX.

In a previous study, Moraru et al. (2012) reported that FQs pollutants in the research area derived entirely from inputs of

chicken manure, which was therefore regarded as the sole input resource. Whenever a single type of manure is used in vegetable planting, FQs concentrations would certainly increase with the application of more manure. In the present study, of the two manure-related factors (chicken manure type

and manure quantity), the former plays a more important role than the latter and there was interaction between the two factors. In practice, when fresh chicken dung is used, the average application required for adequate plant nutrition is always less than 9.0 kg/m², but its high FQs content leads to the highest FQs residue in soil. Although more dry manure (more than 10 kg/m²) is used, it has comparatively low FQs content and therefore causes less FQs residue than fresh manure (Li et al. 2014). Therefore, the type of manure plays a more important role than application quantity. This observation illustrates and confirms that merely reducing the application of manure would be insufficient to significantly reduce FQs pollution in practice.

Crop rotation was identified in our previous research as the major determinant of spatial FQs stratification in soil (Li et al. 2013). In the present study (CFX, EFX, and NFX were calculated as individual dependent variables), the links (from cultivation method to crop rotation and to manure type in Fig. 2) also ascertain the role of crop rotation. Therefore, irrespective of total FQs or the individual FQs, crop rotation is also one of the most important factors in FQs pollution. The relationship is clearly illustrated in Fig. 2. According to our domain knowledge, vegetable planting in the intensive agriculture sector is always regional for convenience of production, sale, and transportation. As a result, intensively reared crops (such as leaf vegetables) are subject to multiple rotation cycles, and widespread use is made of various techniques such as polytunnels (arched plastic greenhouses) to improve yields. Furthermore, solar greenhouses are always employed when growing solanaceous fruits (one of the crop rotations). Therefore, the cultivation method is determined by the crop. Meanwhile, each crop rotation requires different nutrients for their differing crop biomasses, for which the appropriate type and quantity of chicken manure would be calculated. Crop rotation therefore plays a central role in FQs pollution.

Beside the common factors (cultivation method, crop rotations, and chicken manure types), other special factors such as elevation and the level of manure application per greenhouse occur in the BN topology structures of CFX and EFX. It is difficult to clarify their roles. Similarly to elevation, we find that the role is related to the cultivation methods. For instance, most of the polytunnels are located in the southern part of the research area, at higher elevation, whereas the solar greenhouses are mainly located at comparatively low elevation in the north of the research area. However, further study is required to confirm whether this is a coincidence or a logical outcome.

From the results with BNs, it is true to quantify expert beliefs within the models about the dependencies among different variables and to propagate consistently the impact of evidence on the probabilities of uncertain outcomes.

Implication of robust BN for improving risk prediction

While it is possible for the models to automatically learn from the dataset, the samples are often limited and therefore cannot be regarded as representative of the population (Cox 2001). Consequently, a model that has learned from a non-representative sample is unlikely to reliably predict new data or scenarios (Li et al. 2010).

A Bayesian network provides a flexible modeling framework that enables the incorporation of domain knowledge and data learning into the construction of the network structure, thereby providing robust prediction of FQs pollution levels and addressing the issues of sampling bias and model uncertainty. Domain knowledge such as causality or associations between environmental factors and FQs residues can be used to refine the model's structure by adding important associations between nodes representing random variables (such as those not learned, due to sampling bias) or removing inappropriate associations (such as those learned incorrectly). Furthermore, according to domain knowledge, three variables (cultivation methods, crop rotations, and manure type) have important contribution to FQs residues, as also shown in our study case, and should be considered as common predictors (factors) when applying the models to other regions. Other variables, such as elevation and application volume, were only applicable to a specific environment or context, as shown in the learned network structure of our study case.

These special factors are unusual and may be considered according to the specific environments in future applications. The network that was adjusted using domain knowledge had better adaptability than other single learners for new datasets or contexts. Consideration of modeling uncertainty is also essential to optimal prediction. Using Occam's Window, highly qualified models can be selected (Cox 2001) and Bayesian model averaging (Hoeting et al. 1999) can be employed to integrate the predictive outputs from different highly qualified models to provide a robust prediction output. Such improvements are important for assessing exposure and the health effects of FQs pollution. For example, this approach can be used to evaluate the ecotoxic antibiotic effects on plants and soil microorganisms and even potentially adverse impacts such as human allergic/toxic reactions resulting from vegetable intake. Using BMA and Occam's Window, we made a robust prediction of the high-risk probabilities of FQs pollution in 100 selected vegetable greenhouses.

Further, if nodes representing utilities and decision actions are added to the BN, better decision-making functionality for controlling FQs residues can easily be implemented on the basis of the robust predictions provided by our methodology.

Conclusion

The BN analysis illustrates the relationships between FQs pollutants and the factors of cultivation method, crop rotation,

and chicken manure type, and the factors common to some target antibiotics. This demonstrates that the three factors interact and play important roles in FQs pollution. The domain knowledge-based model performs well, and the integrative, robust Bayesian network model achieved the highest estimation accuracy, since it incorporates domain knowledge and model uncertainty. Our encouraging results have implications for the use of BNs to assess FQs pollution and in informing decisions on appropriate remedial measures.

Acknowledgments The authors would like to thank for the support by Promotive Research Fund for Excellent Young and Middle-aged Scientists of Shandong Province (BS2013NY008), National Basic Research Program of China (973 program) (2012CB955500-955503), and the Natural Science Foundation of China (41171344 and 41471376).

References

- Aristilde L, Melis A, Sposito G (2010) Inhibition of photosynthesis by a fluoroquinolone antibiotic. *Environ Sci Technol* 44:1444–1450
- Bidgood TL, Papich MG (2005) Plasma and interstitial fluid pharmacokinetics of enrofloxacin, its metabolite ciprofloxacin, and marbofloxacin after oral administration and a constant rate intravenous infusion in dogs. *J Vet Pharmacol Ther* 28:329–341
- Bouckaert R (1995) Bayesian belief networks from construction to inference. Ph.D. thesis, University of Utrecht
- Breiman L (2001) Random forests. *Mach Learn* 45(1):5–32
- Bugmann G (1998) Normalized gaussian radial basis function networks. *Neurocomputing* 20(1–3):97–110
- Buntine W (1996) A guide to the literature on learning probabilistic networks from data. *IEEE Trans Knowl Data Eng* 8:196–210
- Chow C K, Liu CN (1968) Approximating discrete probability distributions with dependence trees. *IEEE Trans Inf Theory* 14:462–467
- Cooper FG, Herskovits E (1992) A Bayesian method for the induction of probabilistic networks from data. *Mach Learn* 9:309–347
- Cox AL (2001) Risk analysis: foundations, models and methods. Springer, New York
- Friedman N, Geiger D (1997) Bayesian network classifier. *Mach Learn* 29:131–163
- Fu LX, Jiang ZG, Ding HZ, Liu YH (2008) Population pharmacokinetics of enrofloxacin and its active metabolite ciprofloxacin in ill cows. *J Vet Pharmacol Ther* 31:240–245
- Garcia MA, Solans C, Calvo A, Hernandez E, Rey R, Bregante MA, Puig M (2005) Determination of enrofloxacin and its primary metabolite, ciprofloxacin, in pig tissues. Application to residue studies. *Biomed Chromatogr* 19(1):27–31
- Girardi C, Greve J, Lamshoeft M, Fetzer I, Miltner A, Schaeffer A, Kastner M (2011) Biodegradation of ciprofloxacin in water and soil and its effects on the microbial communities. *J Hazard Mater* 198:22–30
- Hoeting AJ, Madigan D, Raftery EA, Volinsky TC (1999) Bayesian model averaging: a tutorial. *Stat Sci* 14:382–417
- Hosmer D, Lemeshow S (2000) Applied logistic regression. John Wiley and Sons, Hoboken
- Hu XG, Zhou QX, Luo Y (2010) Occurrence and source analysis of typical veterinary antibiotics in manure, soil, vegetables and groundwater from organic vegetable bases, northern China. *Environ Pollut* 158:2992–2998
- Jin CX, Chen QY, Sun RL, Zhou QX, Liu JJ (2009) Eco-toxic effects of sulfadiazine sodium, sulfamonomethoxine sodium and enrofloxacin on wheat, Chinese cabbage and tomato. *Ecotoxicology* 18:878–885
- John HG, and Langley P (1995) Estimating continuous distributions in Bayesian classifiers. Eleventh conference on uncertainty in artificial intelligence, San Mateo, 338–345
- Keerthi SS, Shevade SK, Bhattacharyya C, Murthy K RK (2001) Improvements to Platt's SMO algorithm for SVM classifier design. *Neural Comput* 13(3):637–649
- Kuchta SL, Cessna AJ (2009) Lincomycin and spectinomycin concentrations in liquid swine manure and their persistence during simulated manure storage. *Arch Environ Contam Toxicol* 57:1–10
- Kumar K, Gupta SC, Baidoo SK, Chander Y, Rosen CJ (2005) Antibiotic uptake by plants from soil fertilized with animal manure. *J Environ Qual* 34:2082–2085
- Li LF, Wang JF, Leung H, Jiang CS (2010) Assessment of catastrophic risk using Bayesian network constructed from domain knowledge and spatial data. *Risk Anal* 30:1157–1175
- Li LF, Wang JF, Leung H, Zhao SS (2012) A Bayesian method to mine spatial data sets to evaluate the vulnerability of human beings to catastrophic risk. *Risk Anal* 32:1072–1092
- Li XW, Xie YF, Wang JF, Christakos G, Si JL, Zhao HN, Ding YQ, Li J (2013) Influence of planting patterns on fluoroquinolone residues in the soil of an intensive vegetable cultivation area in northern China. *Sci Total Environ* 458:63–69
- Li XW, Xie YF, Li CL, Zhao HN, Zhao H, Wang N, Wang JF (2014) Investigation of residual fluoroquinolones in a soil–vegetable system in an intensive vegetable cultivation area in Northern China. *Sci Total Environ* 468:258–264
- Liu L, Liu YH, Liu CX, Wang Z, Dong J, Zhu GF et al (2013) Potential effect and accumulation of veterinary antibiotics in *Phragmites Australis* under hydroponic conditions. *Ecol Eng* 53:138–143
- Martinez JL (2008) Antibiotics and antibiotic resistance genes in natural environments. *Science* 321:365–367
- Morales-Casique E, Neuman PS, Vesselinov VV (2010) Maximum likelihood Bayesian averaging of airflow models in unsaturated fractured tuff using Occam and variance windows. *Stoch Env Res Risk A* 24:863–880
- Moraru R, Pourcher AM, Jadas-Hecart A, Kempf I, Ziebal C, Kervarrec M et al (2012) Changes in concentrations of fluoroquinolones and of ciprofloxacin-resistant Enterobacteriaceae in chicken feces and manure stored in a heap. *J Environ Qual* 41:754–763
- Parshikov IA, Heinze TM, Moody JD, Freeman JP, Williams AJ, Sutherland JB (2001) The fungus *Pestalotiopsis guepini* as a model for biotransformation of ciprofloxacin and norfloxacin. *Appl Microbiol Biotechnol* 56:474–477
- Quinlan R (1996) Improved use of continuous attributes in C4.5. *J Artif Intell Res* 4:77–90
- Russell S, Norvig P (2010) Artificial intelligence: a modern approach. Prentice Hall, Englewood Cliffs
- Smukler SM, Jackson LE, Murphree L, Yokota R, Koike ST, Smith RF (2008) Transition to large-scale organic vegetable production in the Salinas Valley, California. *Agric Ecosyst Environ* 126:168–188
- Sukul P, Spiteller M (2007) Fluoroquinolone antibiotics in the environment. *Rev Environ Contam Toxicol* 191:131–162
- Turiel E, Martin-Esteban A, Tadeo JL (2006) Multiresidue analysis of quinolones and fluoroquinolones in soil by ultrasonic-assisted extraction in small columns and HPLC-UV. *Anal Chim Acta* 562:30–35
- Van Doorslaer X, Dewulf J, Van Langenhove H, Demeestere K (2014) Fluoroquinolone antibiotics: an emerging class of environmental micropollutants. *Sci Total Environ* 500–501:250–269
- Xiao Y, Chang H, Jia A, Hu JY (2008) Trace analysis of quinolone and fluoroquinolone antibiotics from wastewaters by liquid chromatography-electrospray tandem mass spectrometry. *J Chromatogr A* 1214:100–108
- Zhao L, Dong YH, Wang H (2010) Residues of veterinary antibiotics in manures from feedlot livestock in eight provinces of China. *Sci Total Environ* 408:1069–1075