# An E-commerce System for Mining Distributed Spatial Databases

Aleksandar Lazarevic, Dragoljub Pokrajac, Zoran Obradovic

*Abstract*--**Due to restrictive data access and a lack of appropriate data mining software, spatial information from physically dispersed sites is often not properly exploited in e-commerce. In the proposed distributed spatial knowledge discovery system for e-commerce, a secure centralized server collects proprietary heterogeneous data from subscribed businesses as well as relevant data from public and commercial sources and then integrates knowledge to provide valuable management information to subscribers. Considered knowledge discovery methods include: (1) providing estimated values for unobserved, typically expensive attributes of interest to a particular business; or (2) delivering learned models for generalizing extracted knowledge. An evaluation on large, highly nonlinear simulated data suggests that both approaches can provide profitable, effective and useful management recommendations in spatial e-commerce applications.**

## I. Introduction

In various e-commerce domains involving spatial data (real estate, environmental planning, precision agriculture), participating businesses may increase their economic returns and improve environmental stewardship using knowledge extracted from spatial databases. However, in practice, spatial data is often inherently distributed at multiple sites. Due to security, competition and a lack of appropriate knowledge discovery algorithms, spatial information from such physically dispersed sites is often not properly exploited.

Many large-scale spatial data analysis problems also involve an investigation of relationships among attributes in heterogeneous data sets. Instead of applying global recommendation models across entire spatial data sets, designing an ensemble of local models is preferable to better match site-specific needs thus improving financial benefits [1].

One of the applications that may prosper from novel techniques for analysis of spatial data is precision agriculture aimed at lowering production costs and protecting the environment by controlling the environmental characteristics at a sub-field level [2]. This can be achieved by collecting more and better information and by extracting useful knowledge from data, so the farmers can make the more suitable decisions and thus successfully accomplish their multifaceted goals. This is possible by employing technological advances, such as global positioning systems, combine-mounted yield monitors, and computer controlled variable rate application equipment, that provide an opportunity for improving upon traditional approaches of treating agricultural fields uniformly. Profitability of precision agriculture, the risk of equipment incompatibility and its obsolescence are one of the largest concerns listed by farmers, who are generally interested in this new approach, especially if the costs are modest.

A possible approach towards overcoming all these limitations is developing a distributed spatial knowledge discovery system for precision agriculture. In the proposed system a centralized server provides methods for conversion of protocols and data formats, such that customers have not to be concerned about data incompatibility due to obsolete and non-standardized equipment. The server collects proprietary site-specific spatial data from subscribed businesses as well as relevant data from public and commercial sources and integrates knowledge in order to provide valuable management information to subscribed customers. In general, there are two methods for providing useful recommendation actions. The first method assumes distributed spatial data sets with different sets of attributes. Here, the estimation for unobserved (typically expensive) attributes of interest to a particular business can be made according to similarity among the observed attributes with data from another source where the desired attribute is available. The second method includes constructing models for generalizing knowledge extracted from spatial data and delivering them to subscribed customers. However, sometimes the prediction problem in spatial data sets can be extremely complex since a large number of attributes may influence the target attribute and also significant amounts of noise can exist in data.

Given a number of distributed, both heterogeneous and homogeneous spatial data sets, a profitability evaluation of the proposed methods is discussed in Section 2. Extensive experimental results, reported in Section 3, provide evidence that both methods can be computationally efficient and fairly helpful in developing useful management decisions in precision agriculture and other spatial e-commerce applications.

The authors are with School of Electrical Engineering and Computer Science, Washington State University, Pullman, WA 99164-2752, USA. Zoran Obradovic is also with Center for Information Science and Technology, Temple University, Philadelphia, PA 19122, USA.

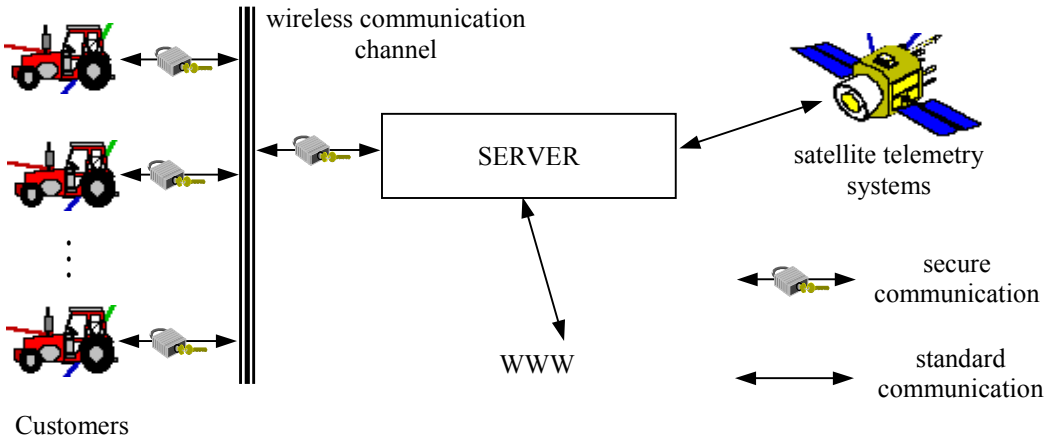



Fig. 1. The scheme of acquiring information to the secure centralized server

## II. METHODOLOGY

In the proposed distributed spatial knowledge discovery system farmers interested in improving their management are subscribed to the centralized server, and they communicate to the server through a wireless communication channel. The system collects two groups of relevant spatial attributes from subscribed businesses. The first group consisting of x and y coordinates and the target attribute is collected by integrating the combine-mounted yield monitors and GPS units. The second group of relevant attributes includes typically expensive soil characteristics which are not necessary available from all farmers. After performing a local data reduction [3] all attributes are transferred through a wireless channel to the centralized server (Fig. 1).

In addition to obtained information from subscribed customers, the server acquires relevant information from other sources available on the World Wide Web (WWW). Publicly available relevant information needed for giving profitable advises include an average temperature, air humidity, precipitation etc., as well the forecasts for those attributes. Commercial information of interest include topographic attributes e.g. slope, elevation, topographic indices, etc. and they are collected through miscellaneous business services, e.g. global positioning systems (GPS), satellite telemetry systems, remote sensing etc. (Fig. 1). The server uses commercial services for acquiring relevant information for all customers, and it shares the costs among all subscribed businesses, hence allowing customers to gain interesting information for less money. In addition, customers are provided with many useful recommendations resulting from integrating knowledge obtained from other subscriber's data.

### *A. Telecommunication subsystem*

A telecommunication sub-system (Fig. 2) consists of the base station located at the site of the central server and mobile stations located on each of harvesting machines that collect crop yield information. Base and mobile stations communicate through a wireless communication channel. A mobile station collects crop yield information, requests communication and after the request is acknowledged, starts with transmission of a data packet from its transmitting buffer (① in Fig. 2). Each data packet contains collected data as well as information for control, error correction and mobile station identification. In the meantime, new-collected data are gathered into an acquisition buffer. The base station receives messages, identifies the sender, and performs error control. If an error in data transmission is detected and cannot be corrected, the base station requests retransmission of an information packet from the corresponding sender by broadcasting the request code along with the sender id (② in Fig 2.). When a wireless channel is not available, a mobile station retries data transmission after a fixed or a random time interval (③ in Fig. 2). A telecommunication subsystem can utilize either of accepted wireless data-transmission multiplex techniques [4]: frequency (FDMA), time-division (TDMA) or code-division multiple access (CDMA). In this paper, we suggest an application of Carrier-Sense Multiple Access/Collision Detection (CSMA/CD) systems that belong to the class of TDMA [5].

In CSMA/CD systems, there is no centralized assignment of the channel to a particular user. Instead, transmitter perceives the channel and starts transmission if there is no signal detected. Due to a propagation delay, another transmitter may broadcast at the same time, when a collision occurs. Once a collision is detected, transmitter retransmits according to one of adopted algorithms. In the simplest case, non-persistent CSMA/CD, which is to be discussed afterward, retransmission occurs after a random time interval. In this paper we will discuss following aspects of an applied telecommunication system:

- the maximal number of mobile stations
- an average number of retransmissions
- the size of an acquisition buffer

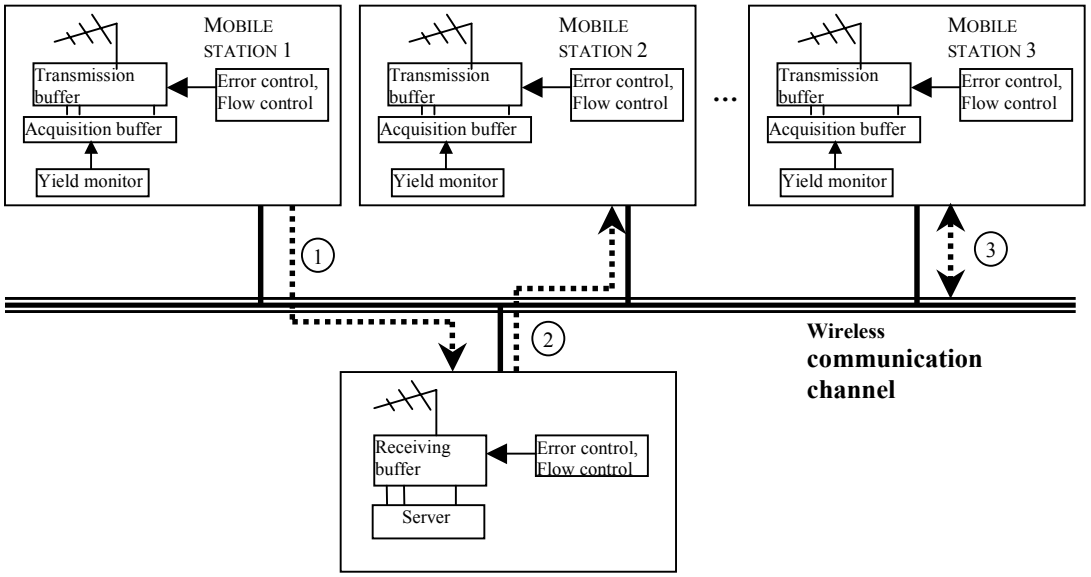



Fig. 2. The scheme of the telecommunication subsystem

*1) The maximal number of mobile stations*

Recall that a yield monitor with global positioning system (GPS) collects one record with the current latitude, longitude and crop yield information each $t_c$ seconds. Let $P$ is the size of transmitting buffer, of which $P'$ bytes are reserved for the transmission of collected records.

The maximal number of mobile station $N_{max}$ can be estimated using the theory of maximal line utilization [5] as:

$$N_{\max} \le \frac{\rho_{\max}}{\lambda m} = \frac{\frac{P'}{P}\frac{t_c f_B}{b}}{1+6.44\frac{2df_B}{cP}} \qquad (1)$$

where $\rho_{max}$ *is* the maximal line utilization, $\lambda$ is customer arrival rate, $m$ is the customer average service time, $d$ is the average distance between a mobile and a base station, $f_B$ is the system byte transmission rate, $b$ is the number of bytes per data record and $c$ is the light speed.

The maximal number of mobile stations increases with the increase of $f_B$, $t_c$ and $P'$ (assuming that the number of control bytes $P$-$P'$ is constant) and decreases with the increment of $b$ and $d$. Typically, sampling rate $t_c$ depends on required yield sampling density and thus cannot be arbitrarily varied. Also, the size $P$ of transmission buffer is limited due to economic reasons. Transmission rate $f_B$ depends on the bandwidth of a wireless channel. On the other hand, the number of bytes per data record depends on the resolution of collected data, while the average distance $d$ depends on various factors, such as the carrier frequency of wireless channel and terrain configuration. For typical values of parameters: $d$ = 20 km, $P$ = 2Kbytes, $f_B$ = 8KBytes/s, $t_c$ = 1s and $b$ = 10 bytes, we obtain small values of propagation delay *0.13ms*, and a high maximal line utilization, $\rho_{max}$ = 0.996. If we choose utilization $\rho$ = 0.8 (to prevent problems that can occur when working near maximal utilization) and assume $P'/P$>0.5 (which is rather pessimistic), we obtain $N_{max}$=320 that satisfies practical requirements for the maximal number of users in an agriculture system

*2) The average number of retransmissions*

The number $J$ of retransmissions due to collisions satisfies geometric distribution [5] with the average number of retransmissions $J_{avg}=1/v$, where $v=Np(1-p)^{N-1}$ is the probability that an attempted transmission occurs without a collision. Parameter $p$ is the probability that in a given moment a particular station occupies the channel. Using an expression for the average time for a successful transmission in the case of non-persistent CSMA/CD [5] the probabilities $v$ and $p$ can be shown to satisfy:

$$\left.\begin{aligned} v &= Np(1-p)^{N-1} \\ p &= \frac{m}{t_P}\left(1+\frac{2df_B}{cP}\left(1+\frac{2}{v}\right)\right) \end{aligned}\right\} \qquad (2)$$

For adopted values of system parameters, after few iterations of (2), one can obtain $p$=0.0025 and $v$=0.196. Hence, the average number of retransmission is $J_{avg}\approx 5$.

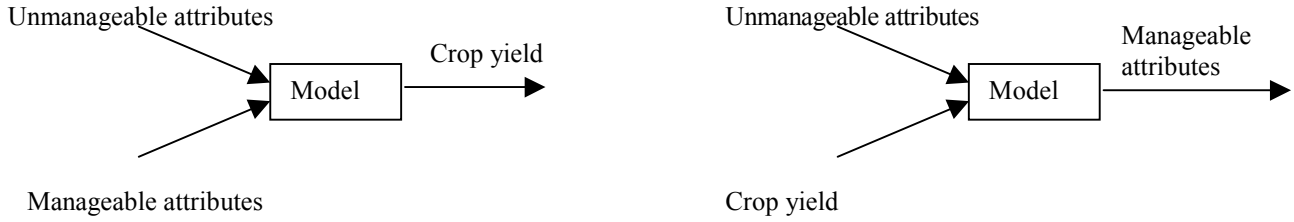
*3) The size of an acquisition buffer*

While re-transmitting, a mobile station acquires new data. To accomplish the quality of service, the probability that data overflows an acquisition buffer must be held within specified boundaries. Given a small probability $\alpha$, we choose the size B of an acquisition buffer such that

$$Prob\left(B \ge \frac{2\tau b}{t_c}J + P'\right) > 1-\alpha \qquad (3)$$

Due to a geometric distribution of $J$, it can be shown that the size of an acquisition buffer must satisfy:

$$B \ge \frac{2\tau b}{t_c}\left\lceil \frac{\log\alpha}{\log(1-v)} \right\rceil + P' \qquad (4)$$




a) Inverse modeling b) Direct modeling

Fig. 3. Basic modeling approaches in agriculture KDD process

Therefore, the size of the acquisition buffer must be larger that the size of the transmission buffer. However, due to $\tau << t_c$, the effect of this acquisition buffer "enlargement" is practically negligible. Even with very high quality requirements ($\alpha$=0.01%), to satisfy (4) it is enough to set $B=P'+1$. Therefore, the acquisition buffer can practically be realized by an addition of one 1-byte shift register to the transmission buffer.

*4) Security aspects*

Businesses in general do not like to share their data due to competitive and other reasons, and therefore one of the most important issues is to verify customer's identity and to provide a confidential and secure communication with the central server.

Customer authentication in the proposed system can be implemented by computing a Message Authentication Code (MAC) [6] as a function of a secret key and the message. This MAC is then appended to the message. Both the customer (sender) and the server (receiver) share the same secret key, where the server uses this key to decide if the message is sent by the customer who claims to have sent it (the only other person with the same secret key).

For communication between customers and the centralized server, subscribed businesses share the same wireless communication channel, and there is an exposure to eavesdropping data transfer on the channel. Therefore, there is also a need to achieve confidentiality of communication, i.e. to pass information between two parties (customer and server) without a third party being able to understand it. Confidentiality can be addressed by encrypting all data packages sent by customers and then decrypting by the server upon receiving the data. For that matter, the customers and the server may use the same keys (private key encryption – DES) or different keys (public key encryption – RSA) [6]. Furthermore, the server should be trustworthy to all customers, in order to be allowed to collect their proprietary relevant information.

## B. Data Processing Subsystem

In the proposed system the server collects a large amount of spatial data from different sources. To maintain and analyze this data there is a need for a server with large secondary storage devices, huge memory capacity, and a high processing speed. When data from all sources is collected at the server site, they are organized into a database with spatial indexing (e.g. R-tree [7]).

Spatial data mining software [8] interfaces this database to extract interesting and novel knowledge from data. Specific objectives include a better understanding of spatial data, discovering relationships between spatial and non-spatial data, construction of spatial knowledge-bases, query optimization and data reorganization in spatial databases. Knowledge extracted from spatial data can consist of characteristic and discriminant rules, prominent structures or clusters, spatial associations and other forms.

Challenges involved in spatial data mining include multiple layers of data, missing attributes and high noise due to a low sensibility of instruments and to spatial interpolation on sparsely collected attributes. To address some of these problems, data is cleaned by removing duplicates, removing outliers and by filtering through a median filter with a specified window size.

## C. Knowledge Discovery Methods

The goal of precision agriculture management is to estimate and perform site-specific crop treatment in order to maximize profit and minimize environmental damage. Through a knowledge discovery (KDD) process, learning algorithms perform data modeling using data sets from different fields in possibly different regions and years. Each data set may contain attributes whose values are not manageable, (e.g. topographic data) as well as these attributes that are manageable (e.g. nutrient concentrations).

Approaches to the modeling in agriculture KDD process supported by our proposed system include a direct and inverse attributes optimization (Fig. 3).

In an inverse modeling, crop yield is modeled based on both unmanageable and manageable attributes. This yield prediction helps farmers to distinguish regions in a field with high and low yield potential and henceforth to adjust an agronomic practice appropriately. A sensitivity analysis of the obtained model, along with techniques of mathematical optimization are used to estimate the optimal concentration of manageable attributes resulting in site-specific fertilizer concentration recommendations.

In a direct modeling, the task of a learning algorithm is to predict one of more manageable attributes using other available attributes and the target attribute (crop yield). This approach provides a direct estimation of a manageable attribute concentration and therefore can help in

determining the optimal treatment for an attribute. Furthermore, it is possible to attempt predicting an attribute whose values are not measured on a particular farm providing relatively cheap nutrient information instead of relaying entirely on an expensive data collection from soil sampling and a subsequent chemical analyses.

The main requirements imposed to learning algorithms employed in precision agriculture are to:

- provide predictions with a sufficiently high generalization,
- allow an user-comprehensible explanation of an observed phenomena,
-discover and exploit spatially similar regions,
-handle noisy data, including sensor and interpolation error and an unexplained yield variance.

The potentials and drawbacks of several knowledge discovery algorithms (ordinary least squares (OLS) linear regression, neural networks, clustering algorithms) are investigated in this paper.

Ordinary least squares (OLS) linear regression is a common method to explain variability of a dependent variable as a linear combination of observed explanatory variables. Weighting coefficients for particular influences are obtained using minimization of a residual error on training data. Linear regression is computationally feasible for large data sets and it provides reasonably robust models of linearly-dependent data. However, when the process to be modeled is highly non-linear, predictions obtained using linear regression are less accurate and there is a need for more sophisticated methods.

Unlike linear regression, neural networks are capable of modeling non-linear dependence in data. The most widely used neural networks are feed-forward multi-layered neural networks [9] (FF-NN). FF-NN consists of several (usually 2) layers of neurons. Each neuron generates its output as a non-linear function of weighted sum of inputs. Inputs for the first layer consist of normalized values of explanatory variables. Outputs of neurons in each layer become inputs of neurons in the subsequent layer. Finally, the output of the last layer becomes a prediction of the response variable. (Fig. 4). On such a way, the output of a FF-NN is a composition of non-linear functions, hence capable of an accurate approximation of a wide class of continuous functions. In practice, neuron non-linearity is usually introduced by a logarithmic sigmoid (as we proceed in this paper) or tangent hyperbolic function. To provide a good generalization, neural networks have to be supplied with an amount of data typically larger than when linear models are learned. Otherwise, we can obtain models specialized to the training data and without capability to explain data variability on previously unseen datasets. Furthermore, a neural network represents a "black box" in which data relation and properties are hard-coded. Therefore, their comprehensibility is often questioned among practitioners. Finally, large amounts of noise and sensor error and the presence of data heterogeneity can dramatically decrease a neural network explanatory power making them to perform in some cases even worse than linear regression models.

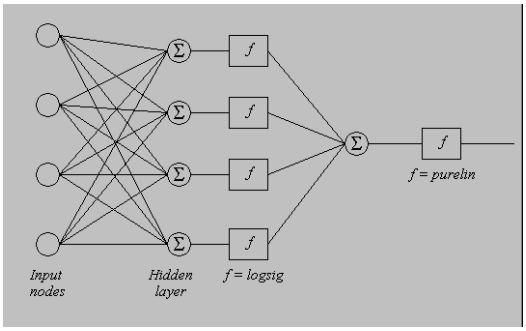



Fig. 4. The architecture of FF-NN model

In order to improve prediction ability when dealing with heterogeneous spatial data, an approach employed in the proposed system is based on identifying spatial regions having similar characteristics using a clustering algorithm. A clustering algorithm is used for partitioning multivariate data into meaningful subgroups (clusters), so that patterns within a cluster are more similar to each other than are patterns belonging to different clusters. Local regression models are built on each of these spatial regions describing the relationship between the spatial data characteristics and the target attribute [10]. Therefore, local models are adapted to specific subsets of the wide range of environments that can exist in spatial data sets even in a small geographic area.

## III. Experimental Results

To illustrate the abilities of the proposed e-commerce system, we performed a series of experiments on simulated data sets. Using simulated data provides a possibility to vary data properties and to determine their impact on the knowledge-discovery process [11]. Experiments were performed on two different collections of simulated data sets.

The first collection consisted of five simulated data sets (fields). For simplicity, each field was a rectangular of $800*800m^2$ with driving attributes influencing the response and corresponding to the relevant soil and topographic attributes in two consecutive years. The soil attributes included levels of Nitrogen, Phosphorus and Potassium, while topographic attributes were Water content and Slope. Each attribute had approximately a normal distribution and statistics (mean value, variance, spatial variability) similar to that of real-life data. Moreover, temporal variability of soil attributes was introduced using AR(1) spatio-temporal model [12]. Piecewise linear models were used to model yield dependence on spatial attributes and AR(1) models to simulate the influence of parameters that vary in time (e.g. weather). Parameters of crop yield models were chosen according to expert knowledge and fertilization guidelines.

Distribution heterogeneity was simulated through a second data collection containing 5 simulated fields with the same attributes that were generated to satisfy the same spatial and temporal properties as in the first collection. However, unlike the first data collection where attributes had approximately normal distributions, in the second collection topographic variables were simulated to be in five clusters, using the technique of feature agglomeration

[12]. Furthermore, instead of using one model for response generation on the entire field, in the second data collection a different data generation process was applied per each cluster.

Knowledge-discovery algorithms were evaluated through the repetitive process of training on a field from one year, and testing on the same data set from the successive simulated year. Prediction accuracy on test data is measured using the coefficient of determination $R^2$ value[1]. The reported prediction accuracy of considered methods was evaluated through 10 trainings of learning models starting from different random initializations of modeling parameters.

### *A. Experiments on homogeneous data*

In experiments with homogenous data both linear (OLS) and non-linear (FF-NN) models were evaluated on the first data collection. We used inverse modeling for prediction of yield and manageable attributes (N,P,K).

In the first experiment, in order to examine generalization capabilities of the proposed methods, models were trained on each of 5 fields and tested on the remaining ones. Depending on a training field, prediction accuracy of linear models expressed through $R^2$ value was within (0.42, 0.54) range. Since yield was simulated using highly nonlinear models, an introduction of nonlinearity in learning algorithms by applying FF-NN resulted in a higher prediction accuracy (average $R^2$ value was within (0.66,0.80) range with standard deviation 0.01-0.02).

Although the simple strategy of building models using data from one field demonstrated promising results, in practice it is often necessary to perform yield prediction on an unseen field by building prediction models on a number of fields. Assuming that the total number of available fields is n, we investigating:

- Training one prediction model on merged (n-1) fields and testing on the remaining one;
- Training one model on each of (n-1) fields and computing prediction accuracy on the test field as an averaged prediction of all (n-1) local models

When one model was trained on merged data, $R^2$ value was in range (0.54,0.59) for OLS and (0.84,0.86) for FF-NN. It is evident that accuracy typically was higher than when the model was trained on only one field, as in previous experiments. Averaging predictions led to a small, but significant increase of FF-NN accuracy ($R^2$ value in range (0.85,0.88)), whereas the accuracy of OLS was the same as in previous case. Finally, we performed a weighted averaging of model predictions. For each point in the test data, weights of particular models were computed according to the similarity of the test point to the distributions of each training set, measured using 4-layer neural networks for learning data distributions on training fields [21]. Due to very similar distributions on all fields, this approach did not lead to an improvement of prediction accuracy. Therefore, for yield prediction on homogeneous fields a simple averaging of prediction models appears to be the most promising technique. This technique is also suitable for distributed databases, where the data sets are physically dispersed, since local models can be trained on the sites where the data are actually stored.

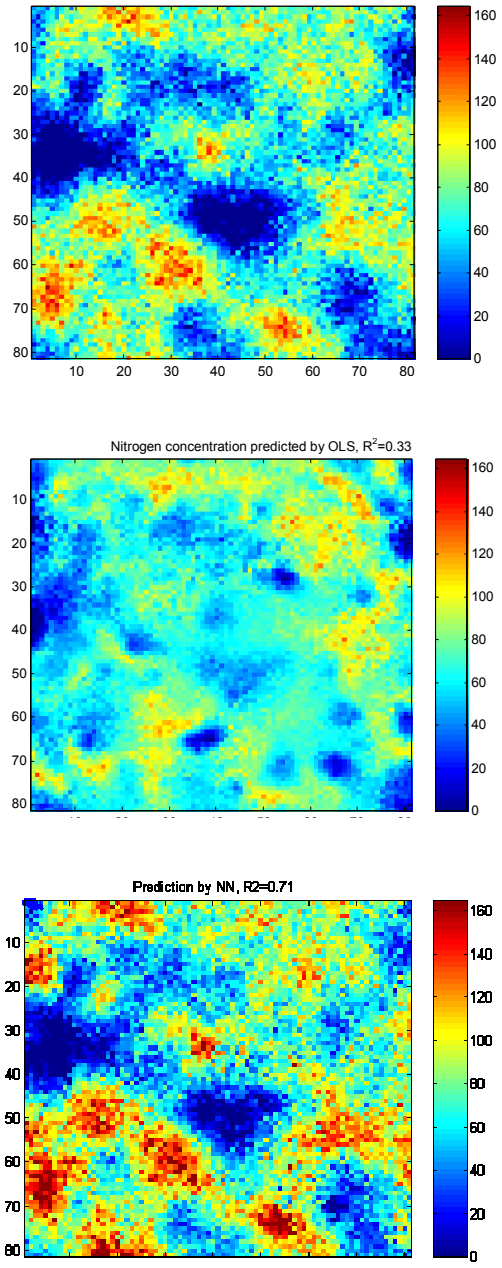


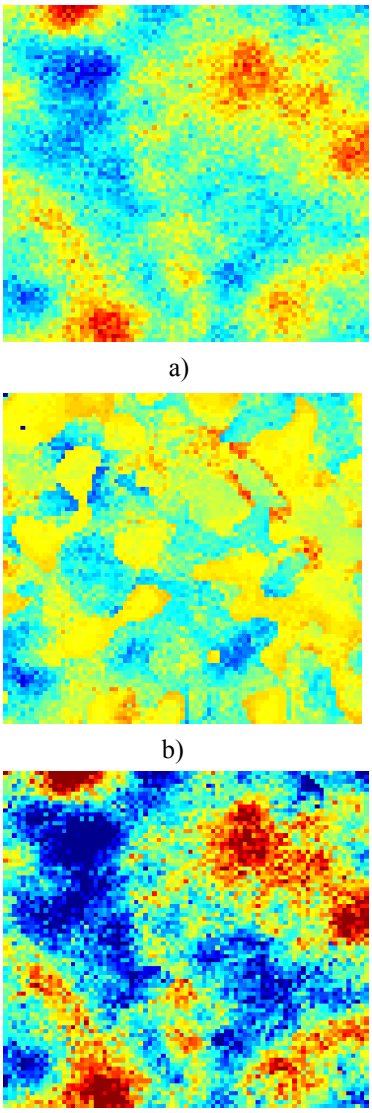

**Fig. 5**. Prediction of Nitrogen concentration on heterogeneous data using direct modeling. a) True concentration b) prediction using OLS and c) prediction using FF-NN.

Next, we performed the prediction of manageable attributes by averaging of local prediction models. Depending on used attributes and a testing field (attributes

[1] Coefficient of determination is equal to $R^2$=1-*average prediction error*/variance(*target variable*). $R^2$ is a measure of the explained variability of the target variable, where larger value is better with 1 corresponding to a perfect prediction and 0 to a trivial mean prediction.

on fields with smaller temporal variability tend to be more predictable using time-lagged data), average prediction accuracy varied in range (0.44,0.74) for FF-NN and (0.13,0.33) for OLS.

As it can be seen from Fig. 5, using OLS results in smoother values of predicted attributes. On the other side, FF-NN result in better prediction ($R^2$=0.71 vs. 0.33 with OLS). Since predicted values of FF-NN resemble high-frequency noise on an image, there might be possible to increase FF-NN prediction accuracy using subsequent 2-D filtering of predicted spatial values.



a)

b)

c)

**Fig. 6**. Prediction of Nitrogen concentration on heterogeneous data using direct modeling. a) True concentration b) Prediction using OLS c) Prediction using FF-NN

*B. Experiments on heterogeneous data*

To investigate performance of the proposed knowledge discovery algorithms on heterogeneous data, we repeated previous experiments on all fields from the second data set. Due to data heterogeneity, yield prediction results obtained using global models were significantly worse than when the same models were applied on homogeneous data. Global FF-NN models trained on one and tested on the remaining fields achieved average $R^2$ value in the range (0.22,0.39) with standard deviation of 0.03-0.04, which was 22-25% worse than for experiments on homogeneous data in the previous section. Due to data heterogeneity, performance of linear models was close to those of non-linear ones: OLS achieved $R^2$ in range (0.19,0.35).

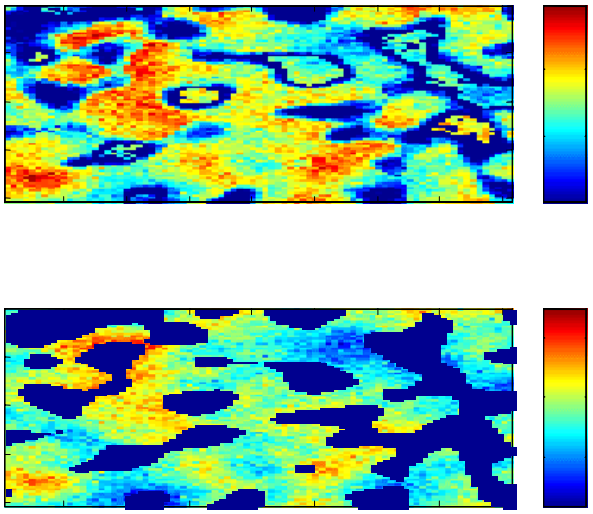
An analogue set of experiments to those for homogeneous data sets, suggests again that prediction achieved by simple averaging of models trained on distinct fields is better than the prediction achieved by applying one model trained on merged data. By averaging, we were able to increase $R^2$ value to (0.17-0.40) range.

In direct modeling of manageable attributes, FF-NN models consistently outperformed OLS. However, due to data heterogeneity, FF-NN was not able to correctly identify regions of low and high attribute values, predicting usually values around the mean of true value, Fig. 6. In contrast, OLS provided a good detection of low- and high-value regions, which can be useful for fertilizer treatment.

To determine benefits of inverse modeling in identifying the optimal concentration of fertilization, we performed a series of experiments on heterogeneous data. Since the outcome of these experiments depends on current market price and local regulations for treatment parameters (e.g. cost of unit of fertilizer, the unit price of crop, the maximal allowed fertilizer concentration) we were not able to provide a general assessment of an overall prediction quality. However, experiments suggested that using inverse modeling can help in successfully discovering regions where profit does not increase if fertilization is performed and in obtaining a fertilization recommendations similar or close to known optima elsewhere (see an example on Fig. 7).

In order to better generalize, the prediction models are constructed for each distribution separately. The total prediction accuracy of the yield was computed as a weighted average of prediction accuracies for learned distributions, where the weights were proportional to the number of instances in each distribution. First, we performed experiments on 5 data sets when all relevant attributes were available for modeling. Since the data sets were simulated through time for two consecutive years, we also tested generalization capabilities of built prediction models in time dimension. Global and local regression models were constructed on each data set from the first year and tested on the same data set from the second year. The experimental results for both model types on all 5 spatial data sets (fields) are shown in Table 1.

TABLE I
THE ACCURACY OF GLOBAL AND LOCAL REGRESSION MODELS THROUGH TIME



b)

Figure 7: Nitrogen fertilization on heterogeneous data. a) the optimal treatment; b) a recommendation obtained using inverse modeling

| Regression Method | Field 1 | Field 2 | Field 3 | Field 4 | Field 5 |
|---|---|---|---|---|---|
| | $R^2 \pm$ std | $R^2 \pm$ std | $R^2 \pm$ std | $R^2 \pm$ std | $R^2 \pm$ std |
| Global | 0.73 ± 0.10 | 0.76 ± 0.15 | 0.78 ± 0.10 | 0.71 ± 0.13 | 0.79 ± 0.06 |
| Local | 0.89 ± 0.04 | 0.91 ± 0.05 | 0.92 ± 0.03 | 0.87 ± 0.04 | 0.92 ± 0.03 |

Although performance of the global models for heterogeneous spatial data sets declined substantially when comparing to using global models on homogeneous data, such an approach can still provide useful capabilities for prediction of the yield in the next year. However, the mixture of local regression models significantly outperformed the global model thus leading to very good generalization abilities.

To simulate the scenario when customers (farmers) do not have access to all relevant soil attributes, and when only topographic attributes are available, we performed experiments on data with different sets of observed soil attributes. In these experiments, Field 1 had attributes corresponding to Nitrogen and Phosphorus, Field 2 had Nitrogen and Potassium, Field 3 had Nitrogen, Phosphorus and Potassium, Field 4 had Phosphorus and Potassium, and Field 5 had only Nitrogen. Generalization capabilities of yield prediction models through time for this scenario are summarized in Table 2.

TABLE II
THE ACCURACY OF GLOBAL AND LOCALPREDICTION MODELS THROUGH TIME WHEN SOME SOIL ATTRIBUTES ARE NOT AVAILABLE

| Regression Method | Field 1 | Field 2 | Field 3 | Field 4 | Field 5 |
|---|---|---|---|---|---|
| | $R^2 \pm$ std | $R^2 \pm$ std | $R^2 \pm$ std | $R^2 \pm$ std | $R^2 \pm$ std |
| Global | 0.48 ± 0.10 | 0.48 ± 0.05 | 0.75 ± 0.12 | 0.48 ± 0.07 | 0.54 ±0.03 |
| Local | 0.63 ± 0.03 | 0.64 ± 0.05 | 0.91 ± 0.06 | 0.65 ± 0.04 | 0.75 ± 0.02 |

Similar to a situation when all relevant attributes are available for modeling, the mixture of local regression models significantly outperformed the method of building global prediction models. However, it is also evident that generalization capabilities of both methods considerably dropped comparing to the previous experiment, where all attributes were available. Hence, it appears that the soil attributes provide a lot of information needed for fairly accurate generalization.

In order to test generalization capabilities of prediction models built on fields with incomplete set of soil attributes, we tested models constructed on those fields on Field 3, where all soil and topographic attributes were available. To further improve achieved prediction accuracy of the yield, we also used simple averaging of models built on different fields. The experimental results for both scenarios, when all relevant attributes were available and when only some of them are obtained, are shown in Table 3.

TABLE III
THE PREDICTION ACCURACY ON FIELD 3 WHEN PREDICTING FROM THE REMAINING FIELDS WITH MISSING ATTRIBUTES

| Used fields in Predicting F3 | Not all attributes available | | All attributes available | |
|---|---|---|---|---|
| | Global | Local | Global | Local |
| F1 | 0.63 ± 0.01 | 0.70± 0.01 | 0.75 ± 0.04 | 0.89 ± 0.01 |
| F2 | 0.70 ± 0.02 | 0.75 0.01 | 0.80 ± 0.03 | 0.90 ± 0.01 |
| F4 | 0.55 ± 0.03 | 0.64 ± 0.01 | 0.77 ± 0.01 | 0.90 ± 0.01 |
| F5 | < 0 | < 0 | < 0 | < 0 |
| F1, F2 | 0.76 ± 0.01 | 0.82 ± 0.01 | 0.84 ± 0.04 | 0.90 ± 0.01 |
| F1, F2, F4 | 0.77 ± 0.02 | 0.84 ± 0.02 | 0.85 ± 0.03 | 0.91 ± 0.01 |
| F1, F2, F4, F5 | 0.66 ± 0.11 | 0.72 ± 0.07 | 0.78 ± 0.04 | 0.81 ± 0.08 |

Analyzing the results from Table 3, it is evident that averaging of models constructed on different fields outperformed generalization from single field models. In addition, the mixture of local regression models was able to improve the prediction accuracy over the global models.

## IV. CONCLUSIONS

A new distributed spatial knowledge discovery system for e-commerce applications is proposed. In the proposed system, the centralized server is collecting proprietary site-specific spatial data and then integrating knowledge in order to provide valuable management information to subscribed customers. An overview of the proposed methodology, with emphasize on telecommunication and security aspects, is provided along with a brief description of the proposed knowledge-discovery techniques. The new approach is successfully applied to several simulated homogeneous and heterogeneous spatial data sets.

Methods for estimating values of unobserved attributes of interest to a particular business as well as the target (crop yield prediction and fertilizer recommendation in our case) were examined. It is shown that a negative influence of distribution heterogeneity on prediction accuracy can be substantially compensated using clustering-based learning algorithms. The extensive experimental results indicate that the proposed system can be computationally efficient and fairly helpful in providing useful recommendations for spatial e-comerce applications.

Although the performed experiments provide evidence that the proposed approaches are suitable for distributed learning in spatial databases, further work is needed to optimize methods for combining models in larger distributed systems.

## REFERENCES


[1] Hergert, W. Pan , Huggins, D., Grove, J., Peck, T., "Adequacy of Current Fertilizer Recomndation for Site-Specific Management," In F. Pierce, *The state of Site-Specific Management for Agriculture*, American Society for agronomy, Crop Science Society of America, Soil Science Society of America,pp. 283-300, 1997.

[2] Hess, J.R., Hoskinson, R.L., "Methods for characterization and analysis of spatial and temporal variability for researching and managing integrated farming systems," *Proceedings of the Third International Conference on Precision Agriculture*, edited by P.C. Robert, R.H. Rust, and W.E. Larson (ASA, CSSA, SSSA, Madison, WI., 1996), pp. 641 – 650.

[3] Vucetic, S. and Obradovic, Z., "Performance Controlled Data Reduction for Knowledge Discovery in Distributed Databases," *Proc. Pacific-Asia Conf. on Knowledge Discovery and Data Mining*, Kyoto, Japan, April 2000.

[4] DeRose, J. F., *The Wireless Data Handbook*, 4th edition, John Wiley and Sons, Inc, 1999.

[5] Schwartz, M. *Telecommunication Networks: Protocols, Modeling, and Analysis*, Addison-Wesley Co, 1986.

[6] Carroll, J., M.: *Computer security*, Boston, Butterworth-Heinemann, 1996.

[7] Gutman, A., "R-trees: a dynamic index structure for spatial searching", *Proceedings of ACM SIGMOD International conference on management of data,* pp.47-57, 1984.

[8] Koperski, K., Adhikary, J., Han, J.,: "Spatial Data Mining: Progress and Challenges", DMKD 1996.

[9] Haykin,S., *Neural networks, a comprehensive foundation,* Prentice-Hall 1999.

[10] Lazarevic, A., Xu, X., Fiez, T. and Obradovic, Z. "Clustering-Regression-Ordering Steps for Knowledge Discovery in Spatial Databases," *Proc. IEEE/INNS Int'l Conf. on Neural Neural Networks*, Washington, D.C., No. 345, session 8.1B., ( 1999)

[11] Pokrajac, D., Obradovic, Z., Fiez, T., "Understanding the influence of noise, sampling density and data distribution on spatial prediction quality through the use of simulated data," *Proc. 14th European Simulation Multiconference (ESM) 2000*, in press.

[12] Pokrajac, D., Fiez, T., Obradovic, Z., "A data generator for evaluating spatial issues in precision agriculture," in review.