REG^2: A Regional Regression Framework for Geo-Referenced Datasets

Oner Ulvi Celepcikay
University of Houston
Department of Computer Science
Houston, TX, 77204-3010
+1-713-743-3356
onerulvi@cs.uh.edu

Christoph F. Eick
University of Houston
Department of Computer Science
Houston, TX, 77204-3010
+1-713-3345
ceick@cs.uh.edu

ABSTRACT
Traditional regression analysis derives global relationships between variables and neglects spatial variations in variables. Hence they lack the ability to systematically discover regional relationships and to build better models that use this regional knowledge to obtain higher prediction accuracies. Since most relationships in spatial datasets are regional, there is a great need for regional regression methods that derive regional regression functions that reflect different spatial characteristics of different regions. This paper proposes a novel regional regression framework that first discovers interesting regions showing strong regional relationships between the dependent and the independent variables, and then builds a prediction model with a regional regression function associated with each region. Interesting regions are identified by running a representative-based clustering algorithm that maximizes an externally plugged in fitness function. In this work, we propose two fitness functions: an R-squared based fitness function and an AIC-based fitness function to handle overfitting better. We evaluate our framework in two case studies; (1) identifying causes of arsenic contamination in Texas water wells and (2) Boston Housing dataset determining spatially varying effects of house properties on house prices. We demonstrated that our framework effectively identifies interesting regions and builds better prediction systems that rely on regional models.

Categories and Subject Descriptors
H.2.8. [**Database Applications**]: *Spatial databases and GIS, Data Mining*

General Terms
Algorithms, Design, and Experimentation

Keywords
Spatial Data Mining, Regression Analysis, Regional Knowledge Discovery, Regional Regression, Clustering.

1. INTRODUCTION
Regression analysis has been extensively used in many scientific fields to discover relationships and dependencies among variables and many variations of regression analysis have been proposed in the literature [2]. In spatial data mining, regression analysis can be used to discover spatial relationships and to build models for prediction. In geo-referenced datasets, most relationships exist at regional level not global level. Since often this type of spatial or regional relationships are not explicitly represented in geo-referenced datasets, one major task in spatial data mining is to develop techniques and methodologies to automate the extraction of interesting and useful regional patterns, and to build better models that reflect the spatially varying characteristics of geo-referenced data. Moreover, capturing regional variation will lead to a deeper understanding of important relationships that are embedded in a spatial dataset. For example, a global linear regression analysis on housing prices in a city would derive coefficients that measures each attribute's contribution to the price of a house. However, coefficients tend to vary from one region to another; for example, the attribute *have_pool* might have a coefficient of 9,000 in a city wide regression analysis which indicates that having a pool adds $9,000 to the house value, when in reality depending on the neighborhood, the pool adds value between $5,000 and $50,000 to the house price, and its contribution differs regionally; e.g. it is much lower for houses in some underdeveloped parts of the city. We claim that understanding these regional variations will not only lead to more accurate prediction models, but will also provide regional background knowledge concerning which attributes have a significant impact on house prices in which regions. Therefore it is desirable to develop methods that capture this spatial variation and extract regional knowledge from spatial datasets. This type of regional knowledge is crucial for domain experts who seek to obtain a deeper understanding of regional variations in relationships among variables in geo-referenced datasets.

Some local regression methods have been proposed in the literature but most use pre-defined region boundaries like zip codes, county limits, or grid structures based on spatial coordinate systems. However, in spatial data mining, regional patterns often do not coincide with predefined geographic boundaries and important spatial relationships might not be discovered due to dilution. For example, proximity to a river might significantly add to the value of the house; but, if census blocks are used as regions, this pattern will not be detected especially if the pre-defined regions contain a lot of houses that are not very close to the river. In general, such relationships can only be discovered by methods that seek for regions on their own that reflects the underlying

structure of the data, e.g. regions that contain houses with similar relationships between dependent and independent variables.

This paper proposes a **Reg**ional **Reg**ression framework called REG^2 (pronounced as REG-squared) that focuses on discovering regional regression functions that are associated with contiguous areas in the subspace of the spatial attributes which we call regions. Figure 1 shows an example of discovered regions along with their regional regression functions and region representatives where the regional regression function for region 10 is given as an example. First, interesting regions are discovered by running a representative-based clustering algorithm that maximizes an externally plugged-in fitness function; next, regional knowledge is extracted from the obtained subspaces (regions). We developed two fitness functions: a R-squared-based fitness function (*RsqFitness*) and an AIC-based (Akaike's Information Criterion) fitness function (*AICFitness*). The two fitness functions are used to guide the search for regions with strong regional linear relationships between the response variable and the independent variables.

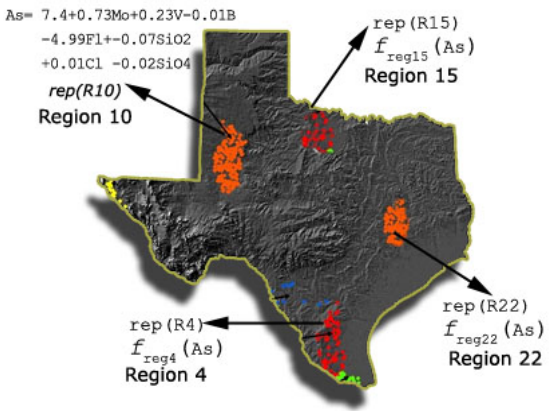

Figure 1. Discovered Regions and Regression Functions

The main contributions of the paper include:

1. A regional regression framework that employs representative-based clustering to discover interesting regions and their associated regional regression functions, without using any predefined region boundaries.

2. An AIC-based and an Rsq-based fitness function to guide the search for regions with highly accurate regression functions.

3. A correlation-based methodology to evaluate the performance of the employed fitness functions and to select fitness function parameters automatically.

4. An experimental evaluation of the framework in a case study that centers on indentifying causes of arsenic contamination in Texas water wells and on Boston Housing dataset determining spatially varying effects of house properties on house prices.

The remainder of the paper is organized as follows: In section 2, we discuss related work. Section 3 provides a detailed discussion of our region discovery framework, the AIC-based fitness function and R-squared based fitness function. Section 4 presents the experimental evaluation, and section 5 concludes the paper.

2. RELATED WORK

2.1 Regression Analysis and AIC

Regression Analysis has been extensively used in many scientific fields to discover linear relationships and dependencies among variables and many variations of regression analysis exists in the literature [2]. OLS is the best known of all regression techniques, which provides a global model of the variable of the interest that needs too be understood or predicted. R^2 is a measure of the extent to which the total variation of the dependent variable is explained by the model. In general, increasing the number of independent variables involved in regression will lead to a higher R^2 value. R^2 alone is not a good measure for goodness of fit of a model since it only deals with the bias of the regression model, and ignores the complexity of a model and its associated variance, and is therefore prone to overfitting. Consequently, several information criteria have been developed that take the complexity of the employed model into consideration, with. Akaike's Information Criterion (AIC)[1] being one of the most popular information criteria to measure the goodness of fit of an estimated statistical model. AIC provides a balance between bias and variance, and is estimated using the following formula:

$$AIC = -2\log L(\hat{\theta}) + 2k \quad (1)$$

where $L(\hat{\theta})$ is the maximized likelihood function, $\hat{\theta}$ is the maximum likelihood estimate of the parameter vector θ under the model and k is the number of the independent parameters of the model. Assuming that the errors are normally distributed, the AIC formula becomes;

$$AIC = 2k + n[\ln(2\pi.SSE/n) + 1] \quad (2)$$

where *SSE* is the Residual Sum of Squares which is called as RSS in some literature and *n* is the number of objects. McQuarrie and Tsai [19] define a special variation of AIC_u for small regions;

$$AIC_u = \ln\frac{SSE}{n-k} + \frac{n+k}{n-k-2} \quad (3)$$

Increasing the number of free parameters to be estimated improves the goodness of fit, regardless of the number of free parameters in the data generating process. Hence AIC not only rewards goodness of fit, but also includes a penalty that is an increasing function of the number of estimated parameters. This penalty discourages overfitting. The preferred model is that with the lowest AIC value. In summary, the AIC methodology attempts to find the model that best explains the data with a minimum of free parameters.

2.2 Regression Trees

Regression Trees are another local statistical prediction model which recursively partitions data into small partitions and then fit a simple model to these small partitions. The early Classification and Regression Tree (CART) algorithm [4] selects the split variable and split value that minimizes the weighted sum of the variances of the target values in the two subsets. The selection of first attribute to split in regression trees dramatically affects the resulted regions and that causes lack of flexibility. Since data is split greedily using a top-down approach, regions in regression trees are rectangular. Regression trees also aim to find local statistics, but our approach is more flexible since it employs an externally plugged-in fitness function to be maximized rather than evaluation variance of splitting on a single attribute like

regression trees employ and also performs wider non-greedy search; moreover, shapes of regions that can be discovered by our approach can be convex polygons, which represent Voroinoi cells whereas regression trees are limited to discovery rectangle shape regions since they discover regions by recursively splitting trees into 2 sub-trees in a top down fashion. Our approach, on the other hand, searches for the optimal set of regions iteratively by modifying region representatives maximizing an external, plug-in fitness function.

2.3 Geographically Weighted Regression

Geographically weighted regression (GWR) [15] is an instance-based, local spatial statistical technique used to analyze spatial non-stationarity. GWR generates maps used in exploring and interpreting spatial non-stationarity. Instead of calibrating a single regression equation, GWR generates a separate regression equation for a set of observation points that are usually determined using a grid structure. Each equation is calibrated using a different weighting of the observations contained in the data set, based on the proximity of observations to observation points. Each GWR equation may be expressed as:

$$y_i = \beta_0(u_i, v_i) + \sum_k \beta_k(u_i, v_i) x_{ik} + \varepsilon_i \quad (4)$$

where (u_i, v_i) denotes the coordinates of the i^{th} point and $\beta_k(u_i, v_i)$ is a realization of continuous function $\beta_k(u, v)$ at point i [21]. Using OLS, the parameters for a linear regression model can be obtained by solving:

$$\beta = (X^T X)^{-1} X^T Y \quad (5)$$

Similarly, the parameter estimates for GWR may be solved using a weighting scheme:

$$\beta(u_i, v_i) = (X^T W(u_i, v_i) X)^{-1} X^T W(u_i, v_i) y \quad (6)$$

The weight assigned to each observation is based on a distance decay function (w_{ij}) centered on observation *i*. Recently, GWR has been used in many research works for investigating a variety of topical areas, including climatology [5], urban poverty [18], environmental justice [20], and the ecological inference problem [7].

GWR is similar to our approach especially in aiming to capture the spatial variance of attributes over space which in GWR is called spatial non-stationarity. Our approach, on the other hand, does not use any observation points, predefined grid structures but it discovers polygon-shaped regions. In general, GWR assumes autocorrelation and will have problems for spatial datasets where patterns change sharply. For example, if we have two neighboring regions that are characterized by very different regression functions, GWR will have significant errors near the boundary of the two regions, because it uses multiple regression functions of nearby observation points and not a single region-specific regression function as does our approach. In general, in our approach a regression function is assigned to each region where coefficient estimates are similar, reflecting the existence of similar pattern of dependency between the dependent and independent variable in a particular region.

2.4 Cluster-Wise Regression (CLR)

The cluster-wise regression technique incorporates cluster analysis into the OLS regression analysis. The simplest cluster-wise regression is a 2-cluster linear regression, which was introduced by Spath [25, 26, 27], and was further developed by other researchers. In summary, instead of using the classical homogeneity or separation criterion, cluster-wise regression is based on the accuracy of a linear regression model associated to each cluster using the sum of squared prediction errors for each object in a cluster. This technique has many applications, due to its being well suited to market segmentation and product pricing [21, 16].

The mathematical formulation of CLR is:

$$Min_e = \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} (y_i - \sum_{j=1}^{m} b_{jk} x_{ij} + b_{0k})^2 \quad (7)$$

given;

$$\sum_{k=1}^{K} z_{ik} = 1 \quad \forall i = 1...n \text{ and}$$

$$z_{ik} \in \{0,1\} \forall i = 1....n \forall k = 1...K.$$

where x_{ij} and y_i are respectively the value of the variable j and the value of the dependent variable for the observation i. b_{jk} is the j_{th} regression coefficients for the cluster k and z_{ij} is a binary variable that equals 1 if and only if the observation i belongs to cluster k. n is the number of observations, m the number of independent variables considered and K the number of clusters. The main objective of CLR is to minimize the sum of squared prediction error for each object using the equation of the cluster that the object belongs to. Our approach is similar but our framework tries to minimize AIC of each region rather than only the prediction error and capability. Moreover, our approach searches for the optimal number of regions rather that assuming that the correct number of regions is known in advance and supports arbitrary, plug-in fitness functions which provide flexibility and extensibility. More importantly, as demonstrated in [6] by Brusco et al., CLR has tremendous potential for overfitting since minimizing the sum of the error sums of squares for the within-cluster regression models makes no effort to distinguish the error explained by the within-cluster regression models from the error explained by the clustering process.

3. METHODOLOGY

We now introduce the components of our regional regression framework, shown in Figure 2. The framework discovers interesting regions by running a representative-based clustering algorithm that maximizes an externally plugged in fitness function. Regional representatives and regional regression functions are associated with regions to support prediction .

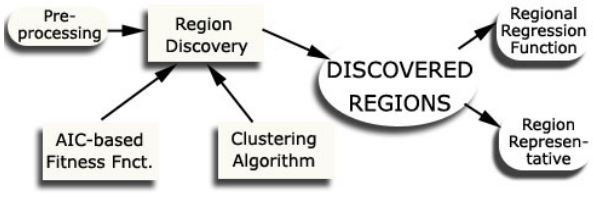

Figure 2. Regional Regression (REG^2) Framework

3.1 Region Discovery Framework

We employ the region discovery framework that was proposed in [13, 14]. The objective of region discovery is to find interesting places in spatial datasets—regions occupying contiguous areas in the spatial subspace. In this work, we extend this framework to extract regional regression functions. The framework incorporates domain knowledge into domain-specific plug-in fitness functions that are maximized by the clustering algorithm. The framework employs a reward-based evaluation scheme to evaluate the quality of the discovered regions. Given a set of regions $R=\{r_1,...,r_k\}$ with respect to a spatial dataset $O=\{o_1,...,o_n\}$, the fitness of R is defined as the sum of the rewards obtained from each region r_j *(j = 1,... ,k)*:

$$q(R) = sum[\psi(r,j)] = \sum_{j=1}^{k} i(r_j) * size(r_j)^{\beta} \quad (8)$$

where $i(r_j)$ is the interestingness of the region r_j—a quantity based on domain interest, reflecting the degree to which the region is "newsworthy." Fitness functions are the core components in the framework, as they capture a domain expert's notion of interestingness. The framework seeks for a set of regions *R* such that the sum of rewards over all of its constituent regions is maximized. In general, the parameter β controls how much premium is put on region size. The size $(r_j)^{\beta}$ component in *q(R),* (β≥1) increases the value of the fitness nonlinearly with respect to the number of objects in the region r_j. $\psi(r,j)$ is the reward of the region and region reward is proportional to its interestingness, but given two regions with the same value of interestingness, a larger region receives a higher reward to reflect a preference given to larger regions. Rewarding region size non-linearly ensures merging neighboring regions whose coefficient estimates are similar reflecting existence of similar pattern of spatial variance of attributes within each region.

3.2 CLEVER Algorithm

We employ the CLEVER [13] clustering algorithm to find interesting regions in the experimental evaluation. CLEVER is a representative-based clustering algorithm that forms clusters by assigning objects to the closest cluster representative and seeks for the optimal set of representatives with respect to q(R). The algorithm starts with a randomly created set of representatives and employs randomized hill climbing by sampling ***s*** neighbors of the current clustering solution as long as new clustering solutions improve the fitness value. To battle premature convergence, the algorithm employs re-sampling: if none of the ***s*** neighbors improves the fitness value, then ***t*** more solutions are sampled before the algorithm terminates. After CLEVER terminates region representatives and their associated regression functions are returned.

3.3 R-squared Fitness Function (RsqFitness)

The natural question in assessing the estimated model is: How well does it fit the data? In our framework we need a fitness function that will be optimized by the clustering algorithm that reflects the main characteristic of good regression models: high accuracy. In regression analysis one of the most popular evaluation measures is R-squared (R^2). Our approach uses the following R^2-based interestingness measure:

Definition 1: (Rsq-based Interestingness – $i_{Rsq}(r)$)

$$i_{Rsq}(r) = \begin{cases} if\ n \geq MinRegSize & 1 - \dfrac{SSE}{SST} \\ if\ n < MinRegSize & 0 \end{cases} \quad (9)$$

R^2 mathematically is equal to 1-SSE/SST. SSE is Sum of Squares of Residuals or Errors and SST is Total Sum of Squares and they are defined as:

$$SSE = \sum_{i-1}^{n} (y_i - \hat{y}_i)^2 \text{ and } SST = \sum_{i-1}^{n} (y_i - \bar{y})^2$$

The *RsqFitness* function then becomes:

$$q_{Rsq}(R) = \sum_{j=1}^{k} i_{Rsq}(r_j) * size(r_j)^{\beta} \quad (10)$$

MinRegonSize is a controlling parameter to battle the tendency towards having very small size regions with maximal variance in regression analysis so that we do not end up with regions with a few objects that have very high R^2 values. The experiments conducted using the R-sq-based fitness function suggest that using R-sq alone frequently leads to lower prediction accuracies on unseen example due to overfitting. We need a better model selection criterion to balance the tradeoff between bias and the variance. Therefore we developed another fitness function that is based on an information criterion, namely AIC.

3.4 AIC-based Fitness Function (AICFitness)

Akaike's Information Criterion (AIC) is one of the most commonly used measures of goodness of an estimated model. We prefer to use AIC because it takes model complexity into consideration; moreover, we believe AIC takes the number of observations more effectively into consideration; there are many variations of AIC including AICu which is used for small size data which we believe is good fit for small size regions. Since lower AIC value indicates a better model and our framework tries to maximize the fitness function, we will use 1/AIC as the interestingness measure. We use AICu if the number of object is less than a predefined threshold *th*, as defined in equation (3). Our work employs the interestingness measure in definition 2 to assess the strength of relationships between the dependent variable and the independent variables in a region *r (also see Section 2.1 for further explanation)*:

Definition 2: (AIC-based Interestingness – $i_{AIC}(r)$)

$$i_{AIC}(r) = \begin{cases} if\ n \geq th & \dfrac{1}{2k + n[\ln(2\pi . SSE/n) + 1]} \\ if\ n < th & \dfrac{1}{\ln \dfrac{SSE}{n-k} + \dfrac{n+k}{n-k-2}} \end{cases} \quad (11)$$

AICFitness function then becomes:

$$q_{AIC}(R) = \sum_{j=1}^{k} i_{AIC}(r_j) * size(r_j)^{\beta} \quad (12)$$

Our fitness function repeatedly applies Regression analysis during the search for the optimal set of regions, minimizing the AIC value in that region. Having an externally plugged in AIC-based fitness function enables the clustering algorithm to probe for the

optimal partitioning and encourages the merging of two regions that exhibit structural similarities.

3.5 The Prediction Schema

The value of the dependent variable for an object is determines as follows: first, the model finds the closest region representative using a 1-nearest neighbor query and then using the regression function associated with this region it predicts the value of the dependent variable. The pseudo-code of this schema is illustrated in figure 3. B_0r is the regression intercept value for the region r and $\beta_j r$ is the regression equation slope coefficient for the independent variable j in region r. *SSE_TE* is the Residual Sum of Squares of testing-set and *SSE_TR* is the Residual Sum of Squares of training-set.

```
    for each object in new_set
{
    -find closest region representative;
    -retrieve the regression function coefficients associated
    with the region (β0r and βjr for all j's);
    -estimate the predicted value of dependent variable (ŷ) by
    using these coefficients;
    -using observed value of dependent variable(y) and
    predicted value (ŷ) estimate SSE_TE(REG2);
    -do same using global model SSE_TE(GL)
}
- output the regional and global RSS (SSE) for new_set
```

Figure 3. The pseudo-code for prediction schema

4. EXPERIMENTAL EVALUATION

4.1 Objectives and Design of the Experiments

One important objective of experimental evaluation is to evaluate the benefits of employing representative-based clustering to discover regional regression functions. Another important objective is to compare the performances of various fitness functions, such as *AICFitness* and *RsqFItness* functions with respect to prediction accuracy. In Section 4.2, a correlation-based methodology for fitness function parameter selection is introduced. We also assess how the regions found using our approach compare with regions that are selected randomly.

In order to evaluate the true performance of our methodologies and to make fair comparisons, our experimental benchmark is composed of four steps. First, we apply OLS regression on the global data. Then we use our framework to discover regions and regional regression functions using both R-sq-based fitness function and AIC-based fitness function. We compare the R^2 value of newly discovered regions vs. global R^2 and also, more importantly, the Residual Sum of Squares (SSE) of new results vs. global SSE to determine the total accuracy improvement.

One could argue that any method that divides data into sub-regions increases the R^2 value and reduces the SSE value, since smaller numbers of objects are involved. This is true to some extent but to show that the regions discovered by our framework are significantly better than randomly selected regions, we also run experiments where regions are randomly discovered without using any fitness functions. This step is repeated five times and the average of the SSE is taken to be fair to the random region discovery. So each dataset was evaluated using the following four benchmarks for each parameter setting: 1) Global Regression, 2) Random Regions, 3) R-sq Fitness Function, 4) AIC-based fitness function. We have used 5-fold cross validation to determine prediction accuracy in the experiments. The SSE values presented in tables or figures represent the values from all 5 folds.

4.2 Parameter Selection Methodology

Utilizing AIC addresses overfitting within a region, but we still must deal with "global overfitting," which can be attributed to having too many regions in regional regression model. The fitness function parameter β is used to control the number of regions to be discovered and thus overall model complexity. The challenge is to find a good value for β for a given dataset that strikes the right balance between underfitting and overfitting for a given dataset. By choosing larger β values model complexity can be penalized, if this desirable for the particular application. This raises the question how we can determine good values for β and other fitness function parameters which is the subject of the remainder of this section. In order to find best parameter combinations we have developed a parameter selection methodology which is based on giving preference for fitness functions that demonstrate a high negative correlation between region reward and the region prediction error on unseen data. N-fold cross validation is used to assess the prediction error on training and test data. Below, we describe in how this correlation (and others involving training accuracy) are computed in detail:

Let

- ψ be the region reward function

- CRV the set of training-set test-set pairs to be used in n-fold cross-validation

- O(r) be the set of objects belonging to region r

- SSE_TE(r) be the test-set sum of squares error of r

- SSE_TR(r) be the training-set sum of squares error of r

- |r| be the number of training set objects belonging to r

FOR EACH (training-set, test-set)-pair in CRV DO

1. *Compute Regions R using training-set;*
2. *FOR EACH region r∈R DO*
   a. *determine number of objects of the test-set that belong to r;*
   b. *determine SSE error for training set and test set objects in r*
   c. *STORE(ψ(r),SSE_TE(r)*|r|) & TS-OBJECTS(r)|,SSE_TR(r))*
3. *Compute correlations between the stored entries in the table*

Basically we determine for each test set of each folding which objects belong to which region, and then we determine the prediction error for those objects using SSE[1] and compute the correlation[2] between regional rewards and regional prediction errors which are expected to be negative since higher rewarded-regions are expected to have lower errors. Table 1 and Table 2 report average correlations between regional rewards and regional predictions errors using different values of β.

[1] Some normalization have to performed to cope with size discrepancies between training set objects belonging to a region (which determine rewards) and test set objects belonging to a region (which determine the error on unseen example).

[2] Additionally, we compute correlations between regional training accuracy and regional testing accuracy and region rewards.

Table 1. Reward & Prediction Error Correlations in Arsenic

Arsenic Dataset	Beta Values			
AICFitness Function	**1.01**	**1.05**	**1.1**	**1.25**
Corr(Reward, SSE_TR)	-0.287	-0.224	-0.265	-0.307
Corr(Reward, SSE_TE)	-0.189	-0.112	-0.211	-0.165

Table 2. Reward & Prediction Error Correlations - Housing

Boston Housing Data	Beta Values			
AICFitness Function	**1.01**	**1.03**	**1.1**	**1.7**
Corr(Reward, SSE_TR)	-0.392	-0.258	-0.48	-0.502
Corr(Reward, SSE_TE)	-0.2	-0.206	-0.377	-0.329

The correlation estimates suggest that there is a negative correlation among regional reward assigned to a region and its prediction error which is an indication of the capability of our framework to identify highly correlated regions that capture and minimize the spatial variation among attributes. Table 3 lists experiments that will be discussed in the remainder of the paper, whose parameters have been selected using the previously described parameter selection methodology. The datasets used in the experiments are both real datasets and are explained next.

Table 3. Final set of experiments and parameters used

Common parameters	*Beta Values(β)*	*Fitness Function*	*Dataset*
Exp# 1	1.01	RsqFitness	Arsenic
Exp# 2	1.03	RsqFitness	Arsenic
Exp# 3	1.05	RsqFitness	Arsenic
Exp# 4	1.1	RsqFitness	Arsenic
Exp# 5	1.01	AICFitness	Arsenic
Exp# 6	1.03	AICFitness	Arsenic
Exp# 7	1.05	AICFitnes	Arsenic
Exp# 8	1.1	RsqFitness	Arsenic
Exp# 9	1.03	Both	Boston Housing
Exp# 10	1.05	Both	Boston Housing
Exp# 11	1.1	Both	Boston Housing
Exp# 12	1.7	Both	Boston Housing

4.3 A Real World Case Study: Texas Water Wells Arsenic Project

Arsenic is a deadly poison, and long-term exposure to even very low arsenic concentrations can cause cancer [28]. Therefore, it is extremely crucial to understand factors that cause high arsenic concentrations to occur. In particular, we are interested in identifying other attributes that contribute significantly to the variance of arsenic concentration. Datasets used in the experiments were created using the Texas Water Department Ground Water Database [28] that samples Texas water wells regularly. The datasets were generated by cleaning out duplicate, missing and inconsistent variables and aggregating the arsenic amount when multiple samples exist.

Our dataset has 3 spatial and 10 non-spatial attributes. Longitude, Latitude and Aqufier ID are the spatial attributes and Arsenic(As), Molybdenum(M), Vanadium(V), Boron(B), Fluoride(F), Silica(SiO2), Chloride(Cl), Sulfate(SiO4) are 8 of the non-spatial attributes which are chemical concentrations. The other 2 non-spatial attributes are Total Dissolved Solids (TDS) and Well Depth (WD). The dataset has 1,653 objects.

4.4 Boston Housing Data-Corrected

We also evaluated our framework using the corrected version of Boston Housing Data which contains 506 census tracts of Boston from the 1970 census. We use the corrected version since this version includes additional spatial information including Longitude and Latitude. The dataset was taken from the StatLib library[27] maintained at Carnegie Mellon University.

We used *MEDV* – the median value of homes in USD 1000's as the target (dependent) variable and 8 of other variables as independent variables. These variables include *CRIM*-crime rate, *NOX*-nitric oxides concentration, *RM*- average number of rooms, *AGE*-proportion of owner-occupied units built prior to 1940, *RAD*-index of accessibility to radial highways, *TAX*-property tax rate, *PTRATIO*-pupil-teacher ratio, *B*-black proportion population, and *LSTAT*- percentage of lower status of the population. We omitted some less important attributes from this regression.

4.5 Arsenic Dataset Results

4.5.1 SSE Improvements

The total Sum of Square of Residuals (SSE) for the global model, Rsq fitness, AIC fitness and the model with randomly discovered regions is shown in figure 4. This SSE values are the Sum of Square of Residuals estimated from the training data.

As shown in the figure, both models with R-sq fitness function and AIC-based fitness function reduce SSE significantly. For example, where global SSE is 76,134 the model with RsqFitness model reduces it to 31,578 and the model with AIC-fitness reduces it to 26,974. SSE that the model using randomly discovered regions produces is 52,716 which still is an improvement over global model as expected since lower number of objects are involved but both our models outperforms Random model as well. In order to illustrate the SSE improvement better, the improvements in percentages are shown in table 4.

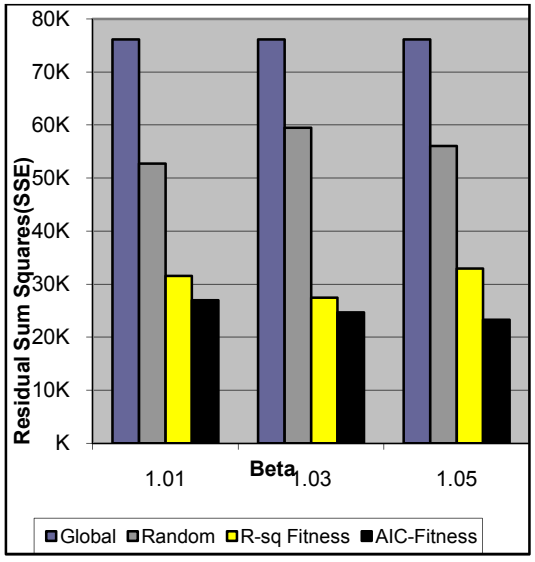

Figure 4. SSE values of four models

Table 4. SSE Improvements over Global Model

Beta Values	*Random*	*R-sq Fitness*	*AIC-Fitness*
1.01	31%	59%	65%
1.03	22%	64%	68%
1.05	26%	57%	69%
1.1	32%	37%	69%

The results show that using randomly selected regions slightly improves the accuracy which is expected due to involvement of less objects. But both the models generated by our framework reduce prediction error significantly. RsqFitness reduces the error by a percentage of high 50s and AICFitness model reduces by almost 70% and produces the best results. The comparison of these three models is sketched in figure 5 for comparison.

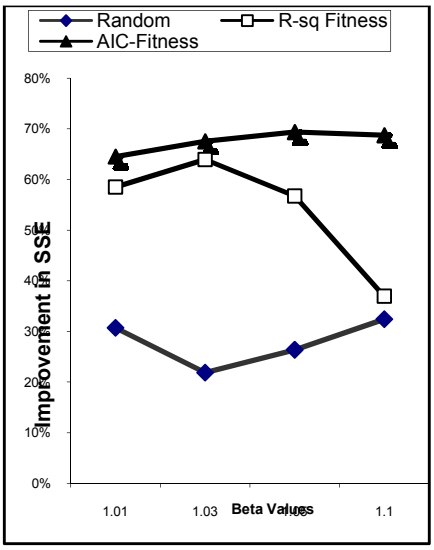

Figure 5. SSE improvements over Global Regression

4.5.2 R^2 and AIC value improvements

The discovered regions illustrate great improvements in R^2 values. R^2 value was 0.6812 for the global data which means 68.12% of the arsenic variation can be explained by other 7 chemical variables for Texas-wide data. The R^2 values for the top 10-ranked regions in experiment 2 are given in Table 5. As can be seen from the table, our framework discovers regions that show high correlation; and in these 10 regions, on average 95% of arsenic variation can be explained using other chemical variables. For example, R^2 value increased from 68% to 98.11% in Region 34 with 195 water wells, which indicates that in this region there exist stronger correlations between arsenic and the chemicals. The R^2 improvements with random region model are also estimated for each experiment. There is some improvement as expected since the regions are much smaller than global, but improvements are much less than the improvements provided by the RsqFitness model. The R^2 values for the top 10-ranked regions using Random region model in experiment 2 are given in Table 6 as an example and they will not be provided for other experiments due to limited space.

Table 5. R^2 value improvements using REG^2

Region	R^2 Value	Size
Texas	**68.12%**	**1655**
Region 34	98.11%	195
Region 27	98.05%	24
Region 1	96.34%	20
Region 35	95.84%	100
Region 20	95.16%	31
Region 3	94.81%	25
Region 22	94.75%	92
Region 20	94.46%	22
Region 33	93.27%	30
Region 6	91.16%	72

Table 6. R^2 value improvements using Random Regions

Region	R^2 Value	Size
Texas	**68.12%**	**1655**
Region 37	96.80%	47
Region 19	89.50%	26
Region 34	84.40%	26
Region 44	82.00%	80
Region 6	79.10%	34
Region 3	78.10%	42
Region 8	75.80%	33
Region 1	73.70%	52
Region 13	71.90%	34
Region 40	71.00%	32

The high R^2 values of regions in Table 5 indicates that our framework successfully discovers regions along with their regional regression functions which represent better model compared to linear regression applied to global data. The average R^2 values of regions in Table 5 is 0.802 which means on average only around 80% of the arsenic variation can be explained by other chemicals. Besides, the high R^2 value of Region 37 and Region 19 dominate the average, otherwise the improvement observed in all other regions can be accepted as insignificant. We now provide the AIC value improvements in Table 7.

Table 7. AIC value improvements in Arsenic

Region	AIC Value	Size
Texas	**11,426**	**1655**
Region 20	3.857	22
Region 10	14.954	21
Region 40	28.021	27
Region 17	32.754	29
Region 35	36.846	43

Again since lower AIC value indicates better model, in fitness function 1/AIC was the fitness value to be maximized. AIC values in discovered regions show great improvement compared to global AIC. Since the AICFitness model provides best accuracy as far as Residual Sum of Squares goes as shown in Table 4, these improvements also indicate the regions of captures spatial variation and minimize the variation within each region (clusters).

4.6 Boston Housing Dataset Results

4.6.1 SSE Improvements

The SSE values of Boston Housing data for the 4 models described previously is shown in figure 6. As shown in the figure, both models with R-sq fitness function and AIC-based fitness function reduce SSE significantly. For example in an experiment with Beta value of 1.01, global SSE is 10,444 and the model with R-sq fitness reduces it to 2,056 and the model with AIC-fitness reduces it to 1,916. SSE that the model using randomly discovered regions produces is 4,165 for these experiments and more than 7,000 in 2 of experiments. SSE improvements in percentages are shown in table 8.

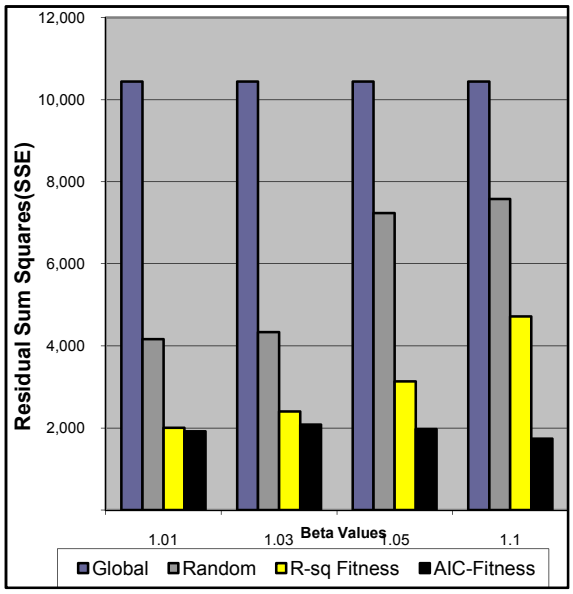

Figure 6. SSE Values of 4 models for Boston Housing Data

Table 8. SSE_TR improvements in Boston Housing Data

Beta Values	Random Regions	R-sq Fitness	AIC-Fitness
1.01	60%	81%	82%
1.03	58%	77%	80%
1.1	31%	70%	81%
1.7	27%	55%	83%

The SSE_TR improvements in the Boston Housing dataset experiments are much better than Arsenic Dataset since this dataset has more correlation and more spatial variance so the search for regions to minimize AIC or maximize Rsq value provides better regions and as a result better prediction accuracy.

As far as SSE_TE is concerned Boston Housing Data also performs better than Arsenic dataset. Due to space limitation all results are not provided but table 9 exemplifies SSE_TE(REG^2) and SSE_TE(GL) calculations which was described in prediction schema in section 3.5. As shown in the table, our framework discovers regions and their regional regression coefficients that perform better prediction compared to the global model. In many regions better prediction is provided and the percentage of such regions is also provided in Table 9. Some regions with very high prediction error reduces the overall prediction accuracy improvement but still there is a significant 27% reduction in testing error which is open for improvement for future work.

Table 9. SSE_TE improvements in Boston Housing Data

β	SSE_TE (GL)	SSE_TE (REG^2)	SSE Improve-ment	% of regions better prediction
1.1	17,182	12,566	27%	72%
1.7	20,028	14,799	26%	65%

4.7 Discussion of Regional Characteristics

Global and regional regression results show that the relationship of the arsenic concentration with other chemical concentrations spatially varies and is not constant over space, which provides a motivation for regional knowledge discovery. In other words, there are significant differences in arsenic concentrations in water wells across various regions in Texas. Some of these differences are found to be due to the varying impact of the independent variables on the arsenic concentration. In order to exemplify this we will compare the global regression model with the regression function of a particular region. The result of the global regression and regional regression of region 10 are shown in Tables 10 and 11, respectively which followed by a discussion these results.

Table 10. Regression Result for Global Data

As	Coef.	Std.Er	t	P>|t|
Mo	0.101	0.0204	4.95	0
V	0.211	0.0048	43.55	0
B	0.0027	0.0003	9.49	0
Fl	-0.6693	0.159	-4.34	0
SiO2	0.0726	0.0115	6.3	0
Cl	0.0008	0.0008	0.97	0.331
SiO4	-0.001	0.0007	-1.87	0.062
const	-1.696	0.4902	-3.46	0.001
R-squared –Value:			68%	
Adjusted R-squared Value			**68%**	

The global OLS regression result suggests that Molybdenum, Vanadium, Boron, and Silica increase the arsenic concentration, but Sulfate and Fluoride decrease it Texas-wide. Moreover, Chloride (Cl) and Sulfate (SiO4) are not significant for global predictors of arsenic concentration but in some regions for example region 10 they are significant. Conversely, Boron, Fluoride, and Silica are globally significant and highly correlated with arsenic, but this is not the case in region 10. This information is crucial to domain experts who seek to determine the controlling

factors for arsenic pollution, as it can help reveal hidden regional patterns and special characteristics for this region. For example, in this region, high arsenic level is highly correlated to high Sulfate and Chloride levels, which is an indication of external factors that play a role in this region, such as a nearby chemical plant or toxic waste. Our framework is able to successfully detect such hidden regional associations between attributes. The global and regional regression results show that the relationship of arsenic concentration with other chemical concentrations spatially varies, and is not constant over space. In addition, there are unexplained differences that are not accounted for by our independent variables, which might be due to external factors, such as toxic waste or the proximity of a chemical plant

Table 11. Regression Result for Region 10

As	Coef.	Std.Er	t	P>|t|
Mo	0.7297	0.2731	2.67	0.013
V	0.234	0.031	7.52	0
B	-0.007	0.004	-1.74	0.094
Fl	-4.996	3.4254	-1.46	0.156
SiO2	-0.071	0.0886	-0.8	0.428
Cl	0.0138	0.0071	2.91	0.066
SiO4	-0.019	0.0142	-3.34	0.192
const	7.3982	4.0134	1.84	0.076
R-squared –Value			95.03%	
Adjusted R-squared Value			**93.73%**	

4.8 Implementation Platform and Efficiency

The components of the framework described in this paper were developed using an open-source, Java-based data mining and machine learning framework called Cougar^2[9], which has been developed by our research group [11]. All experiments were performed on a machine with 1.79 GHz of processor speed and 2GB of memory. The parameter β is the most important factor in determining the run time. The run times of the experiments with respect to the β values used are shown in Figure 7.

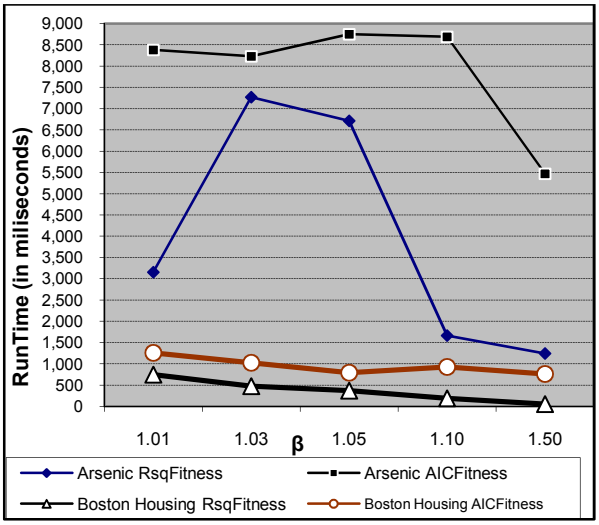

Figure 7. Run Times vs. β Values

In arsenic dataset the β value of 1.05 takes longer than 1.1 or 1.5 but it provides better results as far as SSE improvements and better testing and training set error correlation are concerned. So the time is consumed in exploring better solutions.

The Boston Housing dataset experiments take less time mainly because of two reasons: 1) it has fewer objects 2) more importantly as results provided in section 4 suggest this dataset has more correlation and more spatial variance so the search for regions to minimize AIC or maximize Rsq value takes less time compared to Arsenic data. We observed that more than 80% of the computational resources are allocated for determining regional fitness values when discovering regions. Even though our framework repeatedly applies regression analysis to each explored region combination until no further improvement occurs, it is still efficient compared to approaches in which regression is applied and models were compared that many times using other statistical tools.

5. CONCLUSION

This paper proposes a novel regression framework for spatial datasets, which focuses on discovering regional regression functions that are associated with contiguous areas in the subspace of the spatial attributes that we call "regions." Unlike other research that employs local or global regression, our approach emphasizes regional patterns and provides a comprehensive methodology to help discover them. The proposed framework discovers regions by employing representative-based clustering algorithms that maximize AIC-based or Rsq-based fitness functions; next, regional knowledge (regression functions in our case) is extracted from the obtained regions. In general, different discovered regions capture different relationships between dependent and independent variables. This regional knowledge is crucial for domain experts for understanding the underlying structure of the data.

We also developed a generic correlation-based methodology to evaluate and select fitness functions and parameters of fitness functions for a given dataset. This capability is critical to help deal with overfitting in regional regression for more accurate prediction. Compared to competing approaches, our approach does not assume that regions are *a priori* given, and provides sophisticated clustering techniques to find "good" regions; moreover, we provide a methodology to deal with overfitting and for selecting fitness functions that are suitable for the given data. We claim that none of the competing approaches address both issues in their proposed frameworks.

Our proposed framework was tested and evaluated in a real world case study that analyzes regional correlation patterns among arsenic and other chemical concentrations in Texas water wells. The framework was also tested on the corrected version of the Boston Housing dataset. We demonstrated that our framework can effectively and efficiently identify highly correlated regions, along with the regional regression functions which capture the spatial variation of attributes better than global models and builds better models for prediction.

We plan to develop other versions of *AICFitness* function where the fitness function is scaled to penalize "bad" regions more harshly. We also plan to investigate other information criteria like ICOMP [3] and BIC [23] and regional regression approaches that use validation sets to get a better handle on overfitting.

6. REFERENCES

[1] Akaike, H. Information theory and an extension of the maximum likelihood principle. In B.N. Petrov and F. Csake (eds.), Second International Symposium on Information Theory. Budapest: Akademiai Kiado, 267-281, 1973

[2] Applied Regression Analysis, Wiley Series in Probability and Statistics; Fox, J. (1997).

[3] Bozdogan, H. ICOMP: a new model-selection criterion. In Classification and Related Methods of Data Analysis, H. H. Bock (Ed.), Elsevier Science Publishers, Amsterdam; 599-608, 1988.

[4] Breiman, L., Friedman, J., Olshen, R., Stone, C.: Classification and Regression Trees, Wadsworth, Belmont, CA, 1984.

[5] Brunsdon, C., McClatchey, J. and Unwin, D. (2001). 'Spatial variations in the average rainfall–altitude relationships in Great Britain: an approach using geographically weighted regression', International Journal of Climatology, 21, 455–466.

[6] Brusco, M. J., Cradit, J. D., Steinley, D., & Fox, G. L. (2008). Cautionary remarks on the use of clusterwise regression. Multivariate Behavioral Research, 43 (1), 29-49.

[7] Calvo, C. and Escolar, M. (2003). 'The local voter: a geographically weighted approach to ecological inference', American Journal of Political Science, 47, 189–204.

[8] Choo, J.,Jiamthapthaksin, R., Sheng Chen, C., Celepcikay, O.U., Giusti,C., Eick, C.F.: MOSAIC: A proximity graph approach for agglomerative clustering. The 9th Int'l Conference on Data Warehousing & Knowledge Discovery (DaWaK 2007),Germany, 2007

[9] Cougar^2 Framework, https://cougarsquared.dev.java.net/

[10] Cressie, N.: Statistics for Spatial Data (Revised Edition). New York: Wiley, 1993.

[11] Data Mining and Machine Learning Group, University of Houston, http://www.tlc2.uh.edu/dmmlg

[12] Ding, W., Jiamthapthaksin. R., Parmar R., Jiang D., Stepinski, T., and Eick, C.F., Towards Region Discovery in Spatial Datasets, in Proc. Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD), Osaka, Japan, May 2008.

[13] Eick, C.F., Parmar, R., Ding, W., Stepinski, T., Nicot, J.P.: Finding Regional Co-location Patterns for Sets of Continuous Variables in Spatial Datasets, in Proc. 16th ACM SIGSPATIAL International Conference on Advances in GIS (ACM-GIS), Irvine, California, November 2008.

[14] Eick, C.F., Vaezian, B., Jiang D., , J.: Discovering of interesting regions in spatial data sets using supervised clustering. Proc. of the 10th European Conference on Principles of Data Mining and Knowledge Discovery, Berlin, Germany, 2006.

[15] Fotheringham, A.S., Brunsdon, C. & Charlton, M.: Geographically Weighted Regression: The Analysis of Spatially Varying Relationships. John Wiley, 2002

[16] J.M. Aurifeille. A bio-mimetic clusterwise regression algorithm for consumer segmentation. In Advances in Computational Management, pages 145–163, 1998.

[17] Johnson,R.A.: Applied Multivariate Analysis. Englewood Cliffs, N.J. : Prentice Hall, c1992

[18] Longley, P. A. and Tobon, C. (2004). 'Spatial dependence and heterogeneity in patterns of hardship: an intra-urban analysis', Annals of the Association of American Geographers, 94, 503–519.

[19] McQuarrie A. D., and Tsai, C. (1998), "Regression and Time Series Model Selection", World Scientific Publishing Co. Pte. Ltd., River Edge, NJ.

[20] Mennis, J. and Jordan, L. (2005). 'The distribution of environmental equity: exploring spatial non-stationarity in multivariate models of air toxic releases', Annals of the Association of American Geographers, 95, 249–268.

[21] M.J. Brusco, J. D Cradit, and S. Stahl. A simulated annealing heuristic for a bicriterion partitioning problem in market segmentation. Journal of Marketing Research, 39:99–109, 2002.

[22] Openshaw, S. 1998. "Geographical data mining: key design issues". GeoComputation'99: Proceedings Fourth International Conference on GeoComputation, Mary Washington College, Fredericksburg, Virginia, USA, 25-28 July 1999.

[23] Schwarz, G. (1978). Estimating the dimension of a model. Annals of Statistics, 6, 461-464.

[24] Spath, H., "Clusterwise linear regression," Computing, 22 (4), 367-373, 1979.

[25] Spath, H., "Correction to Algorithm 39: Clusterwise Linear Regression." Computing, 26, 275, 1981.

[26] Spath, H., "Algorithm 48: A Fast Algorithm for Clusterwise Linear Regression." Computing, 29, 175-181,1982.

[27] STATLIB Probability and Statistics Library http://lib.stat.cmu.edu/

[28] Texas Water Development Board, http://www.twdb.state.tx.us/home/index.asp

[29] Woolridge, J.: Econometric Analysis of Cross-Section and Panel Data. MIT Press, 2002, pp. 130, 279, 420-449.