Proceedings of the

1st ACM SIGSPATIAL International Workshop on Data Mining for Geoinformatics (DMG) 2010

offered under the auspices of the

18th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (ACMGIS)

Jessica Lin, Guido Cervone, Nigel Waters
George Mason University

ACM ISBN: 978-1-4503-0430-6.

Contents

Platinum Sponsor

NSF

Gold Sponsor

Google™

Bronze Sponsors

esri®

Microsoft®

DMG 2010 – Organization

Organization Committee

General Co-Chairs

Jessica Lin, *George Mason University, USA*
Guido Cervone, *George Mason University, USA*
Nigel Waters, *George Mason University, USA*

Program Committee Chair

Anthony Stefanidis, *George Mason University, USA*

Program Committee

Peggy Agouris, *George Mason University, USA*
Rafael Amellier, *Stormcenter Communications*
Kirk Borne, *George Mason University, USA*
Wen Chi, *National Chiao Tung University, Taiwan*
Arie Croitu, *University of Alberta, Canada*
Jing (David) Dai, *IBM Research*
Luca Delle Monache, *University Corporation for Atmospheric Research (UCAR)*
Michal Draminski, *Polish Academy of Sciences, Poland*
Shen-Shyang Ho, *University of Maryland*
Andreas Lattner, *University of Frankfurt, Germany*
Germana Manca, *George Mason University, USA*
Amy McGovern, *University of Oklahoma, USA*
Harvey Miller, *University of Utah, USA*
Fabian Moerchen, *Siemens*
Liviu Panait,*Google Inc.*
Brian Rizzo, *University of Mary Washington, USA*
Pang-Ning Tan, *Michigan State University, USA*
Charles Twardy, *OLS Inc.*
Yang Yue,*Wuhan University, China*
Jianting Zhang, *City College of New York, USA*

Additional Reviewers

Sam Blasiak, *George Mason University, USA*
Yuan Li, *George Mason University, USA*
Juan Luo, *George Mason University, USA*

Foreword

Studying, understanding and protecting the earth and its environment are issues of crucial importance for the sustainment and development of our society. Global climate change, severe weather, and catastrophic natural hazards such as volcanic eruptions, earthquakes, hurricanes, floods, etc, require new scientific methodologies for their study. Understanding their governing dynamics and striving towards their timely detection, prediction, and prevention can help protect lives and properties, and minimize economic impact. The field of Geoinformatics focuses on the development of novel scientific algorithms and the implementation of computational methods to provide solutions to pressing earth-related problems.

Recent advances in ground, air- and space-borne sensor technologies have provided scientists from different disciplines an unprecedented access to earth-related data. These developments are quickly leading towards a data-rich but information-poor environment. The rate at which geospatial data are being generated clearly exceeds our ability to organize and analyze them to extract patterns critical for understanding in a timely manner a dynamically changing world. These massive amounts of data require the use of an integrated framework based on Geographic information science (GIS) to address a variety of scientific questions, such as identifying strong patterns, clustering similar data points, detecting anomalies, and abstracting relevant information from sequences of satellite imagery.

The scope of the Data Mining for Geoinformatics (DMG) Workshop is to provide a forum for the exchange of ideas and the establishment of synergistic activities among scientists working in fields such as geographic information science (GIS), data mining, machine learning, geoinformatics, remote sensing, as well as natural hazards, earth and atmospheric sciences. During this half-day event we aim to bring together these scientific communities.

We would like to thank all reviewers and all members of the Program Committee who helped organize the workshop. The Program Committee was constituted of members from the data mining and geoinformatics areas. Each submitted paper was carefully reviewed by at least two reviewers. In addition to the six papers that were accepted as full papers, one paper was accepted as a student paper (equivalent to the PhD showcase that we adapted from ACM SIGSPATIAL GIS 2010). We hope that this is the first of many DMG workshops to come.

Jessica Lin, Guido Cervone, Nigel Waters

Workshop Co-Chairs

A Polygon-based Methodology for Mining Related Spatial Datasets

Sujing Wang[1,2], Chun-Sheng Chen[1], Vadeerat Rinsurongkawong[1], Fatih Akdag[1]
Christoph F. Eick[1]

[1] Department of Computer Science
University of Houston
Houston, TX 77004, USA
Phone: +1 713 743 3345

[2] Department of Computer Science
Lamar University
Beaumont, TX 77710, USA
Phone: +1 409 880 7798

{sujingwa, lyon19, vadeerat, fatihak, ceick}@cs.uh.edu

ABSTRACT

Polygons can serve an important role in the analysis of geo-referenced data as they provide a natural representation for particular types of spatial objects and in that they can be used as models for spatial clusters. This paper claims that polygon analysis is particularly useful for mining related, spatial datasets. A novel methodology for clustering polygons that have been extracted from different spatial datasets is proposed which consists of a meta clustering module that clusters polygons and a summary generation module that creates a final clustering from a polygonal meta clustering based on user preferences. Moreover, a density-based polygon clustering algorithm is introduced. Our methodology is evaluated in a real-world case study involving ozone pollution in Texas; it was able to reveal interesting relationships between different ozone hotspots and interesting associations between ozone hotspots and other meteorological variables.

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, polygon clustering algorithms, mining related datasets, polygon analysis, polygon distance functions

1. INTRODUCTION

Tools that visualize and analyze geo-referenced datasets have gained importance in the last decade, as can be witnessed by the increased popularity of products, such as Google Earth, Microsoft Virtual Earth and ArcGIS. Polygons play an important role in the analysis of geo-referenced data as they provide a natural representation of geographical objects, such as countries, and in that they can be used for the modeling of spatial events, such as air pollution. Moreover, polygons can serve as models for spatial clusters and can model nested and overlapping clusters. Finally, polygons have been studied thoroughly in geometry and they are therefore mathematically well understood; moreover, powerful software libraries are available to manipulate and to analyze and quantify relationships between polygons. Spatial extensions of popular database systems, such as ORACLE and Microsoft SQL Server 2008, support polygon search and polygon manipulation in extended versions of SQL. Surprisingly, past and current data mining research has mostly ignored the capabilities polygon analysis has to offer.

In general, as we will argue in the remainder of the paper, polygon analysis is particularly useful to mine relationships between multiple, related datasets, as it provides a useful tool to analyze discrepancies, progression, change, and emergent events. This work centers on clustering polygons that have been extracted from multiple, related datasets. In particular, a new methodology to mine related, spatial datasets is introduced that consists of a meta clustering module that clusters polygons and a user driven summary generation module that creates a final clustering and other summaries from a polygonal meta clustering. This paper's main contributions include:

- A novel polygon-based methodology for analyzing related, spatial datasets is introduced.
- In contrast to past research, our approach puts a lot of emphasis on the analysis of overlapping polygons that originate from different datasets. Novel distance functions to assess the similarity of overlapping polygons are introduced for this purpose.
- A density-based polygonal meta clustering algorithm is introduced.
- Summary generation algorithms that create the final clustering from meta clusters are proposed. The algorithms rely on a plug-in fitness function to capture user preferences, which is maximized when generating the final cluster.
- The proposed framework is evaluated in a challenging real-world case study involving ozone pollution in the Houston Metropolitan area.

The paper is organized as follows. Section 2 discusses related work. Section 3 introduces distance functions and clustering algorithms for polygons. Section 4 introduces algorithms that generate a final clustering from polygonal meta clusters. Finally, Section 5 evaluates the proposed methodology using ozone pollution case studies and Section 6 summarizes our findings.

ACM GIS '10 , November 2-5, 2010. San Jose, CA, USA.
(c) 2010 ACM ISBN 978-1-4503-0428-3/10/11...$10.00.

2. RELATED WORK

In [1] Joshi et al. propose a DBSCAN-style clustering algorithm for polygons; the algorithm works by replacing point objects in the original DBSCAN algorithm with the polygon objects. In [2], Joshi et al. introduce a dissimilarity function for clustering non-overlapping polygons that considers both spatial and non-spatial attributes. Buchin et al. [12] propose a polygonal time algorithm to compute the Fréchet distance between two polygons. Several papers [21], [20] propose algorithms to compute the Hausdorff distance between polygons. Sander et al. [7] propose GDBSCAN, an algorithm generalizing DBSCAN in two directions: First, generic object neighborhoods are supported instead of distance-based neighborhoods. Second, it proposes other, more complicated measures to define the density of the neighborhood of an object instead of simply counting the number objects within a given radius of a query point.

Zeng et al. [3] propose a meta clustering approach to obtain better clustering results by comparing and selectively combining results of different clustering techniques. In [4] Gionis et al. present clustering aggregation algorithms; the goal is to produce a single clustering that minimizes the total number of disagreements among input clusterings. The proposed algorithms apply the concept of correlation clustering [5]. Caruana et al. [6.] propose a mean to automatically create many diversity clusterings and then measures the distance between the generated clusterings. Next, the hierarchical meta clusters are created. Finally an interactive interface is provided to allow users to choose the most appropriate clustering from meta clusters based on their preferences. In general, [3], [4], and [6] perform meta clustering on a single dataset, whereas our proposed methodology uses meta clustering to analyze relationship between clusters from multiple related datasets.

Our work also relates to correspondence clustering, coupled clustering, and co-clustering which all mine related datasets. Coupled clustering [17] is introduced to discover relationships between two textual datasets by partitioning the datasets into corresponding clusters where each cluster in one dataset is matched with its counterpart in the other dataset. Co-clustering has been successfully used for applications in text mining [18], market-basket data analysis, and bioinformatics [19]. In general, the co-clustering clusters two datasets with different schemas by rearranging the datasets. The objects in two datasets are represented as rows and columns of a dataset. Then, the co-clustering partitions rows and columns of the data matrix and creates clusters which are subsets of the original matrix. Correspondence clustering [9] is introduced by Rinsurongkawong et al. to cluster two or more spatial datasets by maximizing cluster interestingness and correspondence between clusters. Cluster interestingness and correspondence interestingness are captured in plug-in fitness functions and prototype-based clustering algorithms are proposed that cluster multiple datasets in parallel. In conclusion, coupled clustering [17] and co-clustering [18], [19] are not designed for spatial data and they cluster point objects using traditional clustering algorithms. The techniques introduced in correspondence clustering [9] are applicable to point objects in the spatial space whereas this paper focuses on clustering spatial clusters that originate from different, related datasets that are approximated using polygons.

3. DISTANCE FUNCTIONS AND CLUSTERING ALGORITHM FOR POLYGONS

This paper introduces a methodology that uses polygon analysis to mine related datasets, which consists of 3 steps:

1. Collect/Generate polygonal clusters for multiple related datasets
2. Meta cluster polygonal clusters
3. Extract interesting patterns/create summaries from polygonal meta clusters

As far as polygon generation is concerned, our work uses a contouring algorithm called DCONTOUR [14] to generate polygons from continuous density functions or interpolation functions as described in [13], [14]. Moreover, if spatial cluster extensions are given instead, Characteristic shapes [15] and Alpha shapes [16] can be used to wrap polygons around objects that belong to a particular spatial cluster. Both Characteristic shapes and Alpha shapes algorithms create the Delaunay triangulation of the point set and reduce it to a non-convex hull. Polygon generation (Step 1) will not be discussed any further in this paper; this section focuses on Step 2.

3.1 Distance Functions for Polygons

One unique characteristic of our work is that we have to cope with overlapping polygons; past work on polygonal clustering usually assumes that polygons do not overlap and most uses the Hausdorff distance [11] to assess polygon similarity. However, we believe that considering polygon overlap is of critical importance for polygonal clustering of related datasets. Therefore, in addition to the Hausdorff distance, our work proposes two novel distance functions called overlay and hybrid distance functions.

We define a polygon A as a sequence of points A= $p_1,\dots, p_n$, with point p_1 being connected to the point p_n to close the polygon. Moreover, we assume that boundary of the polygon does not cross itself and polygons can have holes inside. Throughout the paper we use the term polygon to refer to such polygons.

3.1.1 Hausdorff Distance

The Hausdorff distance measures the distance between two point sets. It is the maximum distance of a point in any set to the nearest point in the other set. Using the same notation as [11], let A and B be two point sets, the Hausdorff distance $D_{Hausdorff}(A,B)$ for the two sets is defined as:

$$D_{Hausdorff}(A,B) \\ = \max\{max_{a\in A}min_{b\in B}d(a,b), max_{b\in B}min_{a\in A}d(a,b)\}$$

where *d(a,b)* is the Euclidean distance between point *a* and point *b*.

In order to use the Hausdorff distance for polygons, we firstly have to determine how to associate a point set with a polygon. One straight forward choice is to define this point set as the points that lie on the boundary of a polygon. However, computing the distance between point sets that consist of unlimited number of points is considerably expensive. An algorithm that solves this problem for trajectories has been proposed by [20] and the same technique can be applied to polygons.

3.1.2 Overlay Distance

The overlay distance measures the distance between two polygons based on their degree of overlap. The overlay distance $D_{Overlay}(A,B)$ between polygons A and B is defined as:

$$D_{Overlay}(A,B) = 1 - \frac{area(Intersection(A,B))}{area(Union(A,B))}$$

where the function area(X) returns the area a polygon X covers. Basically, the overlay distance is the quotient of the size of the intersection of two polygons over the size of the union of the two polygons. The overlay distance is 1 for pairs of polygons that do not overlap at all.

3.1.3 Hybrid Distance

The hybrid distance function uses a linear combination of the Hausdorff distance and the overlay distance. Because the overlay distance between two disjoint polygons is always 1, regardless of the actual location in space, additionally using the Hausdorff distance provides more precise approximations of the distance between polygons. The hybrid distance function is defined as:

$$D_{Hybrid}(A,B) = \left(w \times D_{Overlay}(A,B)\right) + \left((1-w) \times D_{Hausdorff}(A,B)\right)$$

where w is the weight associated with each distance function ($1 \geq w \geq 0$). Due to the fact that our goal is spatial clustering and we are interested in obtaining meta clusters whose polygons overlap a lot, typically much more weight will be associated with the overlay distance function.

3.2 The POLY_SNN Algorithm

The SNN (Shared Nearest Neighbors) algorithm [8] is a density-based clustering algorithm which assesses the similarity between two points using the number of nearest neighbors that they share. SNN clusters data as DBSCAN does, except that the number of shared neighbors is used to access the similarity instead of the Euclidean distance.

Similar to DBSCAN, SNN is able to find clusters of different sizes, shapes, and can cope with noise in the dataset. However, SNN copes better with high dimensional data and responds better to datasets with varying densities.

In SNN, similarity between two points p_1 and p_2 is the number of points they share among their k nearest neighbors as follows:

$$similarity(p1,p2) = size\ of\ (NN(p1) \cap NN(p2))$$

where $NN(p_i)$ is the k nearest neighbors of a point *pi*.

SNN density of a point p is defined as the sum of the similarities between point p and its k nearest neighbors as follows:

$$density(p) = \sum_{i=1}^{k} similarity\ (p,pi)$$

where p_i is point p's ith nearest neighbor.

After assessing the SNN density of each point, SNN algorithm finds the core points (points with high SNN density) and forms the clusters around the core points like DBSCAN. If two core points are similar to each other, then they are placed in the same cluster. All non-core points which are not similar to any core point are identified as noise points. All non-noise and non-core points are assigned to the cluster of the nearest core point.

When using SNN to cluster polygons, we first calculate the distances between all pairs of polygons using the distance functions discussed in section 2. Next, we identify the K nearest neighbors for each polygon. SNN calculates the SNN density of each polygon using the k nearest neighbors list and clusters the polygons around core polygons using the DBSCAN like algorithm described above.

4. CREATING FINAL CLUSTERINGS FROM POLYGONAL META CLUSTERS

Several forms of summaries can be generated from polygonal meta clusters:

1. Signatures for meta clusters that summarize what characteristics all the objects in the same meta clusters share.
2. Discrepancy mining can be used to create knowledge of how the clusters in a particular meta cluster differ from the clusters in another meta cluster.
3. Final clusterings can be created from a meta clustering.

Section 5 gives some examples of summaries with respect to characteristics and discrepancies of ozone hotspot polygons. The remainder of this section will discuss how to create a “good” final clustering from a set of meta clusters.

Although clustering has been studied for more than 40 years, its objectives and how to evaluate different clustering results is still subject to a lot of controversy; moreover, current research, particularly most ensemble clustering research is still relying on the misconception that a universal, optimal clustering of a dataset exists. However, in general, domain experts seek for clusters based on their domain-driven notion of “interestingness” which usually differs from generic characteristics used by clustering algorithms; moreover, for a given dataset there usually are many plausible clusterings whose value really has to be determined by the domain expert. Finally, even for the same domain expert multiple clusterings, e.g. clusterings at different levels of granularity, are of value. A key idea of this work is to collect a large number of frequently overlapping clusters organized in form of meta-clusters; a final clustering is then created from those meta clusters based on a user’s notion of interestingness.

To reflect what was discussed in the previous paragraph, we assume that our final cluster generation algorithms provide plug-in fitness functions that capture a domain expert’s notion of interestingness which are maximized when generating the final clustering. Meta clustering provides an alternative approach to the traditional ensemble clustering by creating a more structured input for generating a final clustering, also reducing algorithm complexity by restricting choices. In this section, we propose algorithms that create a final clustering by selecting at most one cluster from each meta cluster. Moreover, due to the fact that polygons originated from different datasets typically overlap a lot, we provide an option for the user to restrict cluster overlap in the final clustering. More formally, we develop algorithms that create a final clustering from a meta clustering by solving the following optimization problem:

Inputs:

1. A meta clustering $M=\{X_1, \ldots, X_k\}$ —at most one object will be selected from each meta cluster X_i $(i=1,...k)$.

2. The user provides her own individual cluster reward function $Reward_U$ whose values are in [0,∞).
3. A reward threshold θ_U—low reward clusters are not included in the final clustering.
4. A cluster distance threshold θ_d which expresses how much cluster overlap/coverage she likes to tolerate.
5. A cluster distance function *dist*.

Find $Z \subseteq X_1 \cup \ldots \cup X_k$ that maximizes:

$$q(Z) = \sum_{c \epsilon Z} reward_U(c)$$

subject to:

1. $\forall\, x \in Z\ \forall x' \in Z\ \ (x \neq x' \Rightarrow Dist(x,x') > \theta_d)$
2. $\forall\, x \in Z\ (Reward_U(x) > \theta_U)$
3. $\forall\, x \in Z\ \forall x' \in Z\ \ ((x \in X_i \wedge x' \in X_k \wedge x \neq x'\ ) \Rightarrow i \neq k)$

The goal is to maximize the sum of the rewards of clusters that have been selected from meta clusters. Constraint 1 prevents that two clusters that are too close to each other are both included in the final clustering. Constraint 3 makes sure that at most one cluster from each meta cluster is selected.

Assuming that we have n meta clusters each containing an average of m clusters, there are roughly $(m+1)^n$ final clusterings; for each meta cluster we can either select one of its clusters for inclusion or we might decide not to take any cluster of the meta cluster due to violations of constraints 1 and 2. Constraint 2 is easy to handle by removing clusters below threshold from the meta clusters prior to running the final cluster generation algorithm.

Many different algorithms can be developed to solve this optimization problem, three of which we are currently investigating:

- A greedy algorithm: A greedy algorithm that always selects the cluster with the highest reward from the unprocessed meta clusters whose inclusion in the final clustering does not violate constraints 1 and 2. If there are no such clusters left, no more clusters will be added from the remaining meta clusters to the final clustering.
- An anytime backtracking algorithm: An anytime backtracking algorithm that explores the choices in descending order of cluster rewards; every time a new final clustering is obtained, the best solution found so far is potentially updated. If runtime expires, the algorithm reports the best solution found so far.
- An evolutionary computing algorithm that relies on integer chromosomal representations; e.g. (1,2,3,0) represents a solution where cluster 1 is selected from meta clustering 1, cluster 2 from meta cluster 2,…, and no cluster is selected from meta cluster 4. Traditional mutation and crossover operators are used to create new solutions, and a simple repair approach is used to deal with violations of constraint 1.

The greedy algorithm is very fast (O(m×n)) but far from optimal, the backtracking algorithms explore the complete search space ($O(m^n)$) and—if not stopped earlier—finds the optimal solution if n and m are not very large; however, the anytime approach can be used for large values of m and n. Finally, the evolutionary computing algorithm covers a middle ground, providing acceptable solutions that are found in medium runtime.

5. EXPERIMENTAL EVALUATION

5.1 The Ozone Dataset

Recently, it has been reported by the American Lung Association [24] that Houston Metropolitan area is the 7th worst ozone zone in the US. The Texas Commission on Environmental Quality (TCEQ) is a state agency responsible for environmental issues including the monitoring of environmental pollution in the Texas. TCEQ collects hourly ozone concentration data for metropolitan areas across the state and publishes the data on its website [22]. TCEQ uses a network of 44 ozone-monitoring stations in the Houston-Galveston area which covers the geographical region within [-95.8070, -94.7870] longitude and [29.0108, 30.7440] latitude. On all the figures included in this paper, the X axis represents the latitude range from 29 to 30.4. The Y axis represents the longitude range from -95.8 to -94.8. We downloaded the hourly ozone concentration data from TCEQ's website between the timeframe of April 1, 2009 at 0:00 to November 30, 2009 at 23:00. In addition to the ozone concentrations, we also downloaded the meteorology data including average wind speed, average solar radiation, and average outdoor temperature for the same time slots as the ozone measurements.

Basically, we create polygons that capture ozone hotspots for particular time slot; for each time slot we obtain a set of polygons. In particular the polygons were generated as follows: First, we download the ozone concentration monitored by 44 monitoring sites from TCEQ's website. Next, a standard Kriging interpolation method [25] is used to compute the ozone concentrations on 20×27 grids that cover the Houston metropolitan area. Finally, we feed the interpolation function into the DCONTOUR algorithm with a defined threshold to create sets of polygons, describing polygon hotspots—areas in the spatial dataset whose ozone concentration is above the input threshold. Two polygon datasets are created by using two different density thresholds as inputs for DCONTOUR algorithm. The use of the density threshold 180 creates 255 polygons. These polygons represent areas where the average one hour ozone concentration is above 80 ppb (parts per billion). The density threshold 200 generates 162 polygons that have one hour ozone concentration more than 90 ppb. The current EAP ozone standard is based on an eight-hour average measurement. In order to meet the standard, the eight-hour average ozone concentration has to be less than 0.08 ppm (80 ppb). Therefore, we can consider the polygons that we created are areas where the ozone level exceeds the EPA standard in that hour. Our experiments were conducted using the polygon dataset generated by DCONTOUR with threshold equal to 200.

Ozone formation is a complicated chemical reaction. There are several control factors involved:

- Sunlight measured by solar radiation is needed to produce ozone.
- High outdoor temperatures cause the ozone formation reaction to speed up.
- Wind transports ozone pollution from the source point.
- Time of Day: ozone levels can continue to rise all day long on a clear day, and then decrease after sunset.

We evaluate our methodology in two case studies. The goal of the first case study is to verify that our new polygon distance functions and clustering algorithm for geospatial polygons can effectively cluster overlapped spatial polygons originated from different related datasets. By analyzing additional meteorological attributes such as outdoor temperature, solar radiation, wind speed and time of day associated with polygons, we can characterize each cluster and identify interesting patterns associated with these hotspots. To accomplish this goal we cluster all polygons at all time slots for certain threshold as a single pool of clusters.

In the second case study, we are interested to generate final clusterings that capture a domain expert's notation of interestingness by plugging in different reward functions, e.g., possible maximum range of ozone pollution represented by area of polygons. In order to summarize final clusterings, we also compute the statistical results of ozone pollution control variables such as outdoor temperature, solar radiation, and wind speed.

5.2 Case Study 1: Anaylizing Meta Clusters of Ozone Hotspots

An ozone polygon is a hotspot area that has ozone concentration above a certain threshold. In the first case study, we apply the POLY_SNN clustering algorithm to cluster all the polygons in the ozone dataset in order to find clusters of hotspots.

Figure 1 displays the meta clustering result of 30 clusters found by POLY_SNN using the hybrid distance function and the number of nearest neighbors k set to 5. The dataset consists of 162 polygons created by DCONTOUR using density threshold equal to 200 (90 ppb). Out of 162 polygons, 30% of polygons in the dataset are considered outliers by POLY_SNN. Polygons marked by the same color belong to the same cluster.

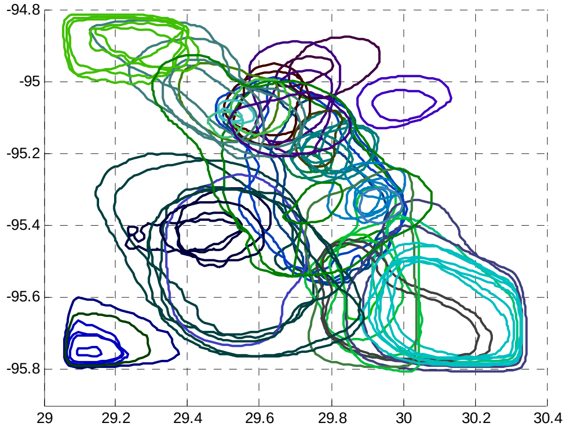


Figure 1. Meta clustering generated by POLY_SNN using the Hybrid distance function.

Table 1. The statistical results of 4 meteorological variables for meta clustering shown in Figure 1

	Mean	Std	Max	Min
Temperature	90.6	5.3	102.8	78.6
Solar Radiation	0.8	0.36	1.4	0.03
Wind Speed	6.1	1.9	15.7	0.3
Time of Day	2:30 pm	1.8	10 am	8 pm

In general, by analyzing the meteorological characteristics of polygons domain experts may find some interesting phenomena that could lead to further scientific investigation. Therefore, we also compute some statistics of 4 meteorological variables involved in ozone formation. Table 1 lists the statistical results of four control factors discussed above associated with the meta clustering in Figure 1.

As expected, meta clustering shown in Figure 1 representing one hour ozone concentration higher than 90 ppd is characterized by high outdoor temperature (average of 90.6 and standard deviation of 5.3) and strong solar radiation (average of 0.80 and standard deviation of 0.36), which usually happens between 1 pm to 4 pm each day. The wind speed affects the range of ozone pollution represented by the size of polygons. Since the standard deviation of the wind speed (1.90) compared with the average wind speed (6.05) is nontrivial, the variation of the size of the polygons is significant in Figure 1.

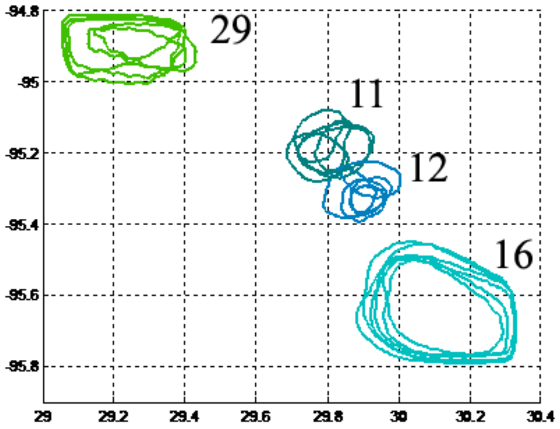


Figure 2. Visualization of 4 meta clusters (ID: 11, 12, 16, and 29) shown in Figure 1.

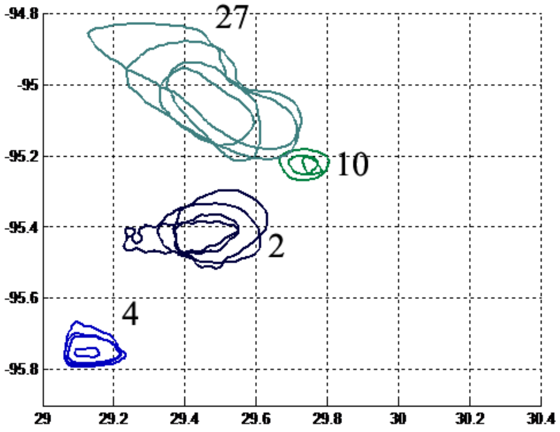


Figure 3. Visualization of 4 meta clusters (ID: 2, 4, 10, and 27) shown in Figure 1.

It is hard to visualize clustering results as polygons overlap a lot as can be seen in Figure 1. Figure 2 and Figure 3 give a picture of eight polygonal meta clusters shown in Figure 1. As expected, the hybrid distance function that employs both overlay distance function and Hausdorff distance function creates clusters of polygons that are similar in shape, size and location. Particularly, since we give more weights to the overlay distance function, the clusters in Figure 2 and Figure 3 are highly overlapped. The clustering results prove that our POLY_SNN clustering algorithm in conjunction with the hybrid distance function can effectively find clusters of overlapping polygons similar in size, shape and location.

Table 2 and Table 3 list the mean and standard deviation of outdoor temperature, solar radiation, wind speed and time of day associated with eight meta clusters in Figure 2 and Figure 3. The solar radiation information related to cluster 2 and 4 are not available from TCEQ's website. Certainly, ozone formation is far more complicated than only considering those four control factors. Our polygon-based methodology has the capability of handling more non-spatial attributes.

Table 2. The statistical results of 4 meteorological variables for 4 meta clusters shown in Figure 2

Meta Cluster Id		11	12	16	29
Temperature	mean	98.83	99.10	90.94	85.48
	std	1.05	2.89	4.26	1.04
Solar Radiation	mean	0.90	0.86	0.70	0.69
	std	0.34	0.0.28	0.28	0.46
Wind Speed	mean	5.16	4.86	5.84	8.34
	std	0.46	0.97	0.93	2.58
Time of Day	mean	2 pm	2 pm	3 pm	12 pm
	std	0.88	1.62	1.63	1.92

Table 3. The Statistical results of 4 meteorological variables for 4 meta clusters shown in Figure 3

Meta Cluster Id		2	4	10	27
Outdoor Temperature	Mean	83.41	88.51	85.95	92.3
	Std	3.81	1.61	2.06	2.86
Solar Radiation	Mean	N/a	n/a	0.65	0.6155
	Std	N/a	n/a	0	0.27
Wind Speed	Mean	6.84	6.15	4.8	6.51
	Std	1.04	0.52	0.79	0.51
Time of Day	Mean	2 pm	1 pm	4 pm	3 pm
	Std	1.70	0.86	0.81	0.83

Based on Table 2, we can see that ozone polygons in clusters 11 and 12 are characterized by very high outdoor temperature (98.83 and 99.10) compared with entire meta clustering (90.6) and strong solar radiation (0.90 and 0.86) compared with entire meta clustering (0.8). The wind speed of cluster 11 and cluster 12 (5.16 and 4.86) are slow compared with entire meta clustering (6.1) so that the average size of the polygons in cluster 11 and cluster 12 are relatively small compared with all other polygons shown in Figure 1. Also, Clusters 11 and 12 are captured around 2 pm. The statistical results associated with Cluster 16 are very close to the entire meta clustering in Table 1. Based on Table 3, cluster 10 has lower outdoor temperature (85.95) compared with entire meta clustering (90.6), lower solar radiation (0.65) compared with entire meta clustering (0.80) and lower wind speed (4.8) compared with entire meta clustering (6.05). The average time of day for cluster 4 is about 4 pm. All those 4 lower meteorological values contribute to smaller polygon sizes inside cluster 4 shown in Figure 3.

5.3 Case Study 2: Final Cluster Generation

The greedy algorithm introduced in section 4 is used to generate the final cluster from polygonal meta clusters shown in Figure 1. We use several reward functions to capture different notations of interestingness of domain experts. The final cluster generated by our model can be used to summarize what characteristics ozone polygons in the same meta clusters share.

The domain experts are usually interested in recognizing the possible maximal range of ozone pollution. The range of ozone pollution represented by polygon area in our model is selected as the first cluster reward function *Reward*$_U$. By selecting different reward threshold and distance threshold, different final clusters could be generated. Figure 4 shows one final cluster using reward threshold 0.04 and Hybrid distance threshold 0.5. There are 5 polygons in the final cluster. A small polygon inside the big dark green polygon is a hole inside the polygon. Our framework allows for polygons with holes inside. Those 5 polygons in Figure 4 clearly capture the dominant ozone hotspots in Houston-Galveston area found in Figure 1.

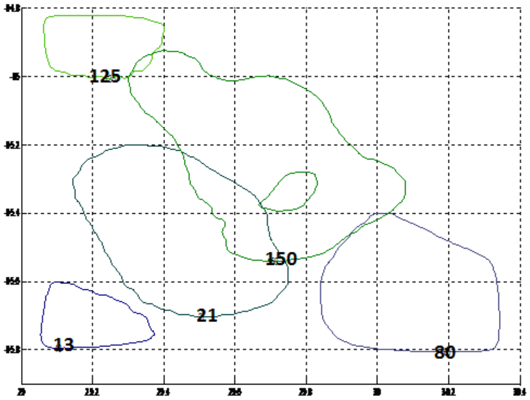


Figure 4. Final cluster for area of polygon reward threshold 0.04 and Hybrid distance threshold 0.5.

Table 4. The mean of 4 meteorological variables for the final cluster shown in Figure 4

Polygon ID	13	21	80	125	150
Outdoor Temperature	79.0	86.35	89.10	84.10	88.87
Solar Radiation	N/A	1.33	1.17	0.13	1.10
Wind Speed	4.50	6.10	6.20	4.90	5.39
Time of Day	6 pm	1 pm	2 pm	2 pm	12 pm

Table 4 shows statistical results of meteorological variables of the final cluster showed in Figure 4. Since the standard deviations of these four variables are relatively small for each polygon, we did not discuss the standard deviation in this section. Based on Table 4, Polygon 21, 80 and 150 covers larger area with higher outdoor temperature, high wind speed and strong solar radiation compared with polygon 12 and 125. Polygon 150 is interesting because it has a hole inside. Further analysis could be done to help understand the formation of holes inside polygons.

The reciprocal of the area of each polygon is used as the second reward function for smaller granularity which may be useful to identify the ozone pollution point source and enable the domain experts to analyze patterns at different levels of granularity. By decreasing either the reward threshold or the distance threshold, we are able to get different final clusters. Figure 5 shows the final cluster with reward threshold set to 10 and distance threshold set to 0.45. There are 14 polygons shown in Figure 5.

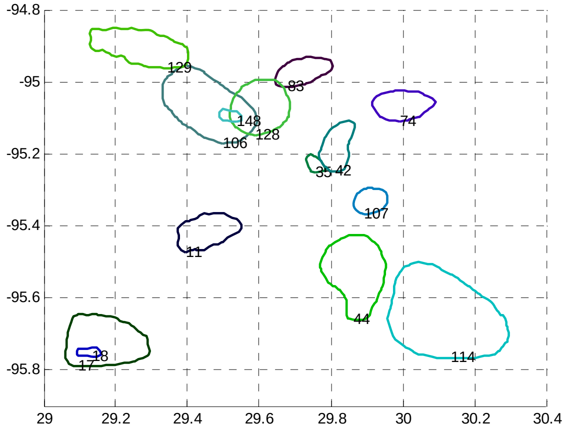


Figure 5. Final cluster for the reciprocal of area reward threshold 10 and Hybrid distance threshold 0.45.

Table 5. The mean of 4 meteorological variables for the final cluster shown in Figure 5

Polygon ID	Outdoor temperature	Solar radiation	Wind speed	Time of day
11	81.4	N/A	6.3	4 pm
17	88.2	N/A	6.0	3 pm
18	N/A	N/A	N/A	4 pm
35	86.3	N/A	6.2	5 pm
42	N/A	N/A	N/A	1 pm
44	N/A	N/A	N/A	3 pm
74	N/A	N/A	N/A	4 pm
83	N/A	N/A	5.9	10 am
106	93.5	0.18	5.9	4 pm
107	94.4	1.21	4.6	11 am
114	94.6	0.63	5.8	4 pm
128	86.4	0.13	5.4	5 pm
129	86.2	1.09	8.8	10 am
148	N/a	N/A	N/A	N/A

Table 5 lists statistical results of four meteorological variables of those 14 polygons shown in Figure 5. Some of the values are not available in the original ozone pollution datasets downloaded from TCEQ website [22]. All of those 14 polygons with relative smaller size occur either before 1 pm or after 4 pm. According to Table 1, the average time of entire meta clustering shown in Figure 1 is 2:30 pm with a standard deviation of 1.8. The time slot from 1 pm to 4 pm everyday is definitely a hotspot for ozone formation which could change the range and the concentration density of ozone pollution significantly. More analysis should be done specially for this time slot.

The outdoor temperatures, wind speed and solar radiation also play a very important role in ozone formation. We use average outdoor temperature associated with each polygon as the third reward function in our model. Figure 6 shows one final cluster with average outdoor temperature threshold set to 90 and distance threshold set to 0.55. There are 5 polygons. The statistical results of meteorological variables are summarized in Table 6. Obviously, all the polygons with high temperatures occur during 2 pm to 4 pm. The lower the wind speed, the smaller the area of the polygon. For example, polygon 67 has the lowest wind speed of 4.1 compared with all the other four polygons in Figure 6.

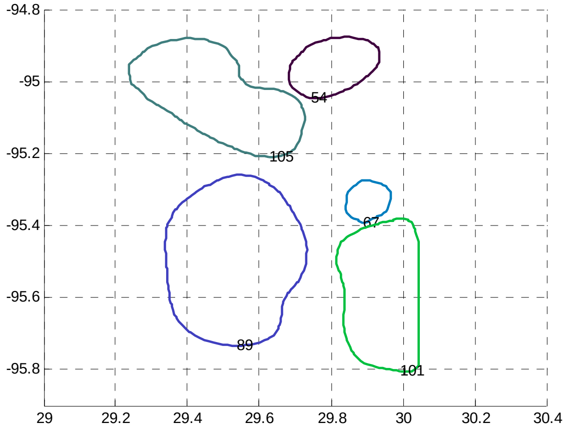


Figure 6. Final cluster for polygon average temperature reward threshold 90 and Hybrid distance threshold 0.55.

Table 6. The mean of 4 meteorological variables of the final cluster shown in Figure 6

Polygon ID	54	67	89	101	105
Outdoor Temperature	100.3	102.8	92.4	99.4	94.5
Solar Radiation	N/A	0.96	0.91	0.70	0.72
Wind Speed	6.0	4.1	8.533	8.2	6.04
Time of day	2 pm	3 pm	3 pm	4 pm	3 pm

6. CONCLUSION

This paper claims that polygon analysis is particularly useful for mining multiple, related spatial datasets. In particular, a novel methodology for clustering polygons that have been extracted from multiple, spatial datasets is proposed which consists of a meta clustering module that clusters the obtained polygons and a summary generation module that extracts patterns and creates summaries from a polygonal meta clustering. In general, this work has the capability to cluster overlapping polygons and use novel distance functions to assess the similarity between polygons. Moreover, a density-based polygonal clustering algorithm called POLY_SNN is proposed by extending SSN. Finally, three algorithms for generating a final cluster from a given meta clustering based on user preferences were discussed. To the best of our knowledge, this is the first paper that proposes a comprehensive methodology that relies on polygon analysis to mine related spatial datasets.

Our methodology is evaluated in a real-world case study involving ozone pollution in the Houston Metropolitan area. It is able to reveal interesting relationships between different ozone hotspots and interesting associations between ozone hotspots and other variables.

7. REFERENCES

[1] Joshi, D., Samal, A. K., and Soh, L.K., 2009. *Density-based clustering of polygons*, in Proc. of IEEE Symposium on Computational Intelligence and Data Mining (Nashville, TN, USA, March 30 - April 02, 2009).

[2] Joshi, D., Samal, A. K., and Soh, L.K., 2009. A dissimilarity function for clustering geospatial polygons, in Proc. of the ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (Seattle, Wa, USA, November 04 -0 6, 2009).

[3] Zeng, Y., Tang, J., Garcia-Frias, J., and Gao, R.G., 2002. *An adaptive meta-clustering approach: combining the information from different clustering results*, in Proc. of IEEE Computer Society Conference on Bioinformatics (Palo Alto,CA, USA., August 14 - 16, 2002).

[4] Gionis, A., Mannila, H., and Tsaparas, P., 2005. *Clustering aggregation*, in Proc. of the International Conference on Data Engineering (Tokyo, Japan , April 05 – 0 8, 2005).

[5] Bansal, N., Blum, A., and Chawla, S., 2002. *Correlation clustering*, in Proc. of Symposium on Foundations of Computer Science (Vancouver, BC, Canada, November 16 – 19, 2002).

[6] Caruana, R., Elhawary, M., Nguyen, N., and Smith, C., 2006. *Meta clustering*, in Proc of IEEE International Conference on Data Mining (Las Vegas,Nevada, USA, June 26 - 29, 2006).

[7] Sander J., Ester M., Kriegel H.-P., and Xu X., 1998. *Density-Based Clustering in Spatial Databases: The Algorithm GDBSCAN and its Applications*, Data Mining and Knowledge Discovery, Vol. 2, No. 2, 1998, pp. 169-194.

[8] Ertoz, L., Steinback, M., and Kumar, V., 2003. *Finding clusters of different sizes, shapes, and density in noisy, high dimensional data*, in Proc. of SIAM International Conference on Data Mining (San Francisco, CA, USA, May 01 - 03, 2003).

[9] Rinsurongkawong, V., Eick, C.F., 2010. *Correspondence clustering: an approach to cluster multiple related datasets*, in Proc. of Asia-Pacific Conference on Knowledge Discovery and Data Mining (Hyderabad, India, June 21 - 24, 2010).

[10] Eick, C.F., Parmar, R., Ding, W., Stepinki, T., and Nicot, J.P., 2008. *Finding regional co-location patterns for sets of continuous variables in spatial datasets*, in Proc. of ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (Irvine, California, USA, November 05 - 07, 2008).

[11] Zhang, Z., Huang, K., and Tan, T., 2006. *Comparison of similarity measures for trajectory clustering in outdoor surveillance scenes*, in Proc. of International Conference on Pattern Recognition (Hong Kong, China, Augest 20 – 24, 2006).

[12] Buchin, K., Buchin, M., and Wenk, C., 2006. *Computing the Fréchet distance between simple polygons in polynomial time*, In Proc. of Symposium on Computational Geometry (Sedona, Arizona, USA, June 05 – 07, 2006).

[13] Rinsurongkawong, V. Chen, C.S., Eick, C. F., and Twa, M., 2010. *Analyzing change in spatial data by utilizing polygon models*, in Proc. of International Conference on Computing for Geospatial Research & Application (Washington, DC, USA, June 21 - 23, 2010).

[14] Chen, C.S., Rinsurongkawong, V., Eick, C.F., Twa, M., 2009. *Change analysis in spatial data by combining contouring algorithms with supervised density functions*, in Proc. Of Asia-Pacific Conference on Knowledge Discovery and Data Mining (Bangkok ,Thailand, April 27 – 30, 2009).

[15] Duckham, M., Kulik, L., Worboys, M., and Galton, A., 2008. *Efficient generation of simple polygons for characterizing the shape of a set of points in the plane*, Pattern Recognition. 41, pp. 3224-3236.

[16] Edelsbrunner, H., Kirkpatrick, D. G., and Seidel, R., 1983. *On the shape of a set of points in the plane*, IEEE Transactions on Information Theory, vol. IT–29, no. 4, pp. 551-558.

[17] Marx, Z., Dagan, I., Buhmann, J.M., and Shamir, E., 2002. *Coupled clustering: a method for detecting structural correspondence*, Journal of Machine Learning Research, pp. 747-780.

[18] Dhillon, I.S., 2001. *Co-clustering documents and words using bipartite spectral graph* partitioning, in Proc. of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (San Francisco, California, USA, August 26 - 29, 2001).

[19] Cheng, Y., Church, C.M., 2000. *Biclustering of Expression Data*, in Proc. of International Conference on Intelligent Systems for Molecular Biology.

[20] Hangouet, J., 1995. *Computing of the Hausdorff distance between plane vector polylines*, in Proc. of Symposium on Computer-Assisted Cartography.

[21] Atallah M.J., Ribeiro, C.C., and Lifschitz, S., 1991. *Computing some distance functions between polygons*, Pattern Recognition, Vol. 24, Issue 8, pp. 775-781, 1991.

[22] Texas Commission on Environmental Quality, http://www.tceq.state.tx.us

[23] MacQueen, J.B.,1967. *Some methods for classification and analysis of multivariate observations*, in Proc. Of Berkeley Symposium on Mathematical Statistics and Probability (Berkeley, CA, USA, June 21 - July 18, 1967).

[24] American Lung Association,

http://www.lungusa.org/

[25] Cressie, N., 1993, *Statistics for spatial data*. New York: Wiley.

HC-DT/SVM: A Tightly Coupled Hybrid Decision Tree and Support Vector Machines Algorithm with Application to Land Cover Change Detections

Jianting Zhang
Department of Computer Science
City College of New York
New York City, NY, 10031

jzhang@cs.ccny.cuny.edu

ABSTRACT

Change detection techniques have been widely used in satellite based environmental monitoring. Multi-date classification is an important change detection technique in remote sensing. In this study, we propose a hybrid algorithm called HC-DT/SVM, that tightly couples a Decision Tree (DT) algorithm and a Support Vector Machine (SVM) algorithm for land cover change detections. We aim at improving the interpretability of the classification results and classification accuracies simultaneously. The hybrid algorithm first constructs a DT classifier using all the training samples and then sends the samples under the ill-classified decision tree branches to a SVM classifier for further training. The ill-classified decision tree branches are linked to the SVM classifier and testing samples are classified jointly by the linked DT and SVM classifiers. Experiments using a dataset that consists of two Landsat TM scenes of southern China region show that the hybrid algorithm can significantly improve the classification accuracies of the classic DT classifier and improve its interpretability at the same time.

Categories and Subject Descriptors

H.2.8 [Database Applications] Data Mining

General Terms

Algorithms, Experimentation, Performance

Keywords

Hybrid Classifier, Decision Tree, SVM, Remote Sensing, Land Cover, Change Detection

1. INTRODUCTION

Change detection from remotely sensed images is a useful technology for detecting changes in large and rapidly changing area and is an important source for environmental monitoring. Many digital change detection techniques have been developed during the past few decades (Singh 1989, Lu et al 2004). The techniques can be grouped into three major categories: map algebra, direct multi-date classification and post-classification comparison. Image (band) differencing might be the most widely used method in the first category. While the techniques in the category are able to provide information on the possible existence of a change and the relative magnitude of the change, they do not identify the nature of the change (Im and Jensen 2005). In contrast, techniques in the later two categories have the capabilities of providing detailed information about the type of land cover change for every pixel and/or polygon under examination (Im and Jensen 2005). While the post-classification comparison based methods are straightforward, they were criticized for relying on the accuracy of the two individual classifications (Singh 1989). In this study, we propose a hybrid algorithm that tightly couples a Decision Tree (DT) algorithm and a Support Vector Machine (SVM) algorithm for land cover change detections that aims at improving the interpretability of the classification results and classification accuracies simultaneously. The proposed approach falls in the multi-date classification category.

Comparisons of different classification algorithms in the multi-date classification category have been extensively studied. For example, Chan et al (2001) compared four classifiers, namely Multi-Layer Perceptron (MLP), Learning Vector Quantization (LVQ), Decision Tree (DT) and Maximum-Likelihood Classifier (MLC). Seto and Liu (2003) compared ARTMAP neural network with MLC and observed that ARTMAP neural network classifiers were more accurate than MLC classifiers. Nemmour and Chibani (2006) has reported that Support SVM generally performed better than a two hidden-layer Artificial Neural Network (ANN) classifier using the standard back propagation rule with respect to classification accuracies. While a certain classifier may have higher classification accuracy for a particular dataset, it is hard to make a conclusion that some classifiers are always better than the rests when multiple criteria are used to evaluate the suitability of algorithms (Chan et al, 2001). Although in reality no classification algorithm can satisfy all evaluation requirements nor be applicable to all studies due to different environmental settings and datasets used (Lu and Weng 2007), hybridizing two or more classifiers with careful design may improve the suitability of classification algorithms for land cover change detections.

Hybrid classifier is a popular concept in classifying remotely sensed data. Various hybrid methods have been proposed since at least early 1990s (Kelly et al 2004). For example, the Iterative Guided Spectral Class Rejection (IGSCR) is a hybrid approach that combines unsupervised clustering and maximum likelihood (Wayman et al 2001) and have been successfully used in a few applications (Kelly et al 2004, Musy et al 2006, Wynne 2007). Techniques that hybridize clustering and classification algorithms for urban change analysis have

ACM SIGSPATIAL DMG Workshop, Nov 2, San Jose, CA, U.S.A.

been successfully applied to the Twin Cities (Minnesota) metropolitan area using multi-date Landsat data (Yuan et al 2005). In addition, the ensemble based techniques, such as bagging or boosting, can also be broadly considered as hybrid classifiers where a same base classifier is applied multiple times and the classifications are combined to generate the final results. In this case, the base classifiers are "hybridized" to themselves. Land cover classifications using boosting (Friedl et al 1999, de Colstoun and Walthall 2006) and bagging (DeFries and Chan 2000, Prasad et al 2006) on decision tree classifiers have been reported. More recently, Nemmour and Chibani (2006) applied multiple support vector machines for land cover change detection where multiple kernels were used to build multiple classifiers and the classification results were combined based on fuzzy integral and attractor dynamic rules. Most of existing hybrid classifiers require training and classifying the samples in a dataset multiple times independently and we term as loosely-coupled hybrid classifiers. The problem with such loosely-coupled hybrid classifiers are that the numbers of training and testing of the hybrid classifiers usually are proportional to the numbers of base classifiers that the hybrid classifiers are based on. Combining classification results from multiple independent classifiers beyond simple majority voting rule requires careful design of schemas of combination (e.g., Liu and Gopal 2004, Huang and Lees 2004) which is a non-trivial task. In addition, while it may be possible to visualize individual base classifiers for better interpretation, it may not be feasible to visualize the hybrid classifiers due to the composition complexities of the base classifiers.

In this study, we propose a new hybrid algorithm that adopts a tightly coupled strategy for base classifiers. The strategy first feeds all the training samples to a fast classifier that uses divide-and-conquer strategy (e.g. decision tree algorithms) and identifies ill-classified components in the divided classification space. The strategy then combines samples fall in the ill-classified components and sends them to a more sophisticated and computationally intensive classifier for further classification. A hybrid classifier that adopts the strategy actually is a chain of two types of base classifiers. One of the advantages of the new type of hybrid classifiers is the capacity to leverage fewer but more significant patterns resulting from the training samples and present them to users immediately in a compact form. For example, delivering decision rules from a resulting decision tree classifier that cover a larger number of training samples with high classification accuracies or few exceptions. In addition, the classifiers that adopt the divide-and-conquer strategy usually can be represented as a tree and can be easily visualized.

As a case study, we have developed a new hybrid algorithm that hybridizes a decision tree classifier and a SVM classifier. Different from previous hybrid techniques that mainly target at classification accuracies, the proposed hybrid algorithm also aims at interpretability of the trained classifier. We choose to hybridize the decision algorithm and the SVM algorithm for two main reasons. First, the decision tree algorithm has been widely used in land cover classification and its advantages, such as no presumption of data distribution and fast in training and execution, have been well recognized (Friedl and Brodley 1997, Friedl et al 2002). More importantly, it has the capabilities of generating human interpretable rules. The decision tree algorithm has been successfully applied to urban change detection as reported in (Chan et al 2001) and (Im and Jensen 2005). Second, recent studies on classifying remote sensing data have consistently reported that SVM classifiers have better classification accuracies than conventional MLC classifiers and ANN based classifiers for both multi-spectral (Huang et al 2002, Pal and Mather 2005) and hyperspectral images (Pal and Mather 2006) which suggests that SVM could be used as an accurate classifier for change detection that involves a large number of bands from multi-date images. Unfortunately, the resulting hyperplane in a SVM classifier is in a high-dimensional space, which makes visualizing SVM classifier for human interpretation very difficult if not impossible.

Following the strategy discussed previously, the proposed hybrid algorithm first applies a decision tree algorithm to the training samples to construct a DT classifier. The samples in the ill-classified branches of the resulting decision tree are used to construct a SVM classifier. The two classifiers are chained together through pointers and used for classification. The proposed approach is motivated by our previous work on devising a successive decision tree algorithm for classifying remotely sensed data where the samples in the ill-classified branches of a previous resulting decision tree are used to construct a successive decision tree (Zhang et al 2007). The rest of the paper is organized as follows. Section 2 introduces the basics of the DT classifier and the SVM classifier and presents the proposed hybrid algorithm. Section 3 provides details of software implementations of the HCC-DT/SVM algorithm. Section 4 is the experiments on the land cover change detections using a pair of TM images at two times in a southern China region. Finally Section 5 is the summary and conclusions.

2. The HC-DT/SVM Algorithm

Before going to the details of the hybrid algorithm, we first briefly introduce the two base classifiers, namely the decision tree classifier and the support vector machines classifier. The hybrid algorithm is then presented as a set of linked procedures.

2.1 The Decision Tree Algorithm

The decision tree method recursively partitions the data space into disjoint sections using impurity measurements (such as information gain and gain ratio). For the sake of simplicity, binary partition of feature space is usually adopted in implementations. Let $f(C_i)$ be the count of class i before the partition and $f(C_i^1)$ and $f(C_i^2)$ be the counts of class i in each of the two partitioned sections based on a partitioning value, respectively. Further let C be the total number of classes, $n=\sum_{i=1}^{C} f(C_i)$, $n_1=\sum_{i=1}^{C} f(C_i^{\,1})$, and $n_2=\sum_{i=1}^{C} f(C_i^{\,2})$, then the information entropy before the partition is defined as $e=-\sum_{i=1}^{C}\frac{f(C_i)}{n}*\log(\frac{f(C_i)}{n})$. Correspondingly the entropies of the two partitions are defined as $e_1=-\sum_{i=1}^{C}\frac{f(C_i^{\,1})}{n_1}*\log(\frac{f(C_i^{\,1})}{n_1})$ and

$e_2 = -\sum_{i=1}^{C} \frac{f(C_i^{\,2})}{n_2} * \log(\frac{f(C_i^{\,2})}{n_2})$, respectively. The overall entropy after the partition is defined as the weighted average of e_1 and e_2, i.e.,

$$entropy\_partition = \frac{n_1}{n} * e_1 + \frac{n_2}{n} * e_2$$

The Information Gain then can be defined as:

$$entropy\_gain = e - entropy\_partition$$

The Gain Ratio is defined as:

$$gain\_ratio = \frac{entropy\_gain}{entropy\_partition}$$

Implementations may choose to use different criteria, such as information gain, gain ratio or their combinations. For example, the J48 module in the WEKA data mining package (Witten and Frank 2000) that implements the popular C4.5 algorithm uses the following procedure to determine the best partitioning attribute (band in remote sensing classification case) and the best partitioning value. First, for each attribute, a set of partitioning values is determined based on the minimum and maximum values of the attribute. Second, each of the partitioning values is used to partition the training samples into two subsets and the information gain and gain ratio are calculated. The partitioning value with the largest gain ratio among the partitioning values whose info gains are above the average is used as the attribute's partitioning value. Third, the process is repeated for all the attributes and the attribute with the largest gain ratio is chosen as the partitioning attribute.

The decision tree classifier adopts a divide-and-conquer strategy and is very fast in training and testing. More importantly, paths from the root to leaf nodes can easily be transformed into decision rules (such as if *a>10* and *b<20* then Class *3*), which is suitable for human interpretation and evaluation. In addition, during the process of selecting partitioning attribute, the algorithm works on an attribute (band) at a time and do not need information from other attributes (bands). Thus band values come from multi-date images do not need rigid radiometric calibration before feeding to the decision tree algorithm. This is a significant advantage of using the algorithm for change detections that involve multi-date images when calibration is difficult.

2.2 The SVM Algorithm

For the sake of simplicity, we only introduce the basic SVM algorithm that handles two classes. For multi-class classification problem, either one against one or one against all strategy can be applied to decompose the multi-class classification problem into multiple two-class classification problems. The SVM classifier we use in this study is the Java version of the LIBSVM package (Chang and Lin 2001) which adopts the one against one class decomposition strategy. An n-class classification problem is decomposed into n*(n-1)/2 two-class classification problems. The results are merged through a majority vote.

For a two-class classification problem, given a set of samples *N* $\{x_i, y_i\}_{i=1}^{N}$, where $x_i \in \mathbf{R}^n$ is the *i*-th sample and $y_i \in \{-1, +1\}$ is the label of the sample, the SVM algorithm aims at finding a linear hyperplane that separate the data in a transformed space, i.e.,

$$y_i[w^T \phi(x_i) + b] \geq 1, i = 1..N$$

where function ϕ(x) is a mapping from the original space to a high dimensional space. In case of such separating hyperplane does not exist, a so called slack variable ξ_i is introduced such that

$$\begin{cases} y_i[w^T \phi(x_i) + b] \geq 1 - \xi_i, i = 1..N \\ \xi_i \geq 0 \end{cases} \tag{1}$$

SVM adopts the structural risk minimization principle and the risk bound is minimized by solving the following minimization problem:

$$\min_{w,\xi} J(w, \xi) = \frac{1}{2} w^T w + c \sum_{i=1}^{N} \xi_i \tag{2}$$

subjected to (1). To minimize (2), a Lagrangian function can be constructed as

$$L(w, b, \xi, \alpha, \beta) = J(w, \xi) - \sum_{i=1}^{N} \alpha_i \{y_i[w^T \phi(x_i) + b] - 1 + \xi_i\} - \sum_{i=1}^{N} \beta_i \xi_i \tag{3}$$

where $\alpha_i \geq 0,\ \beta_i \geq 0\ (i = 1, \ldots, N)$ are the Lagrangian multipliers of (2). Function (3) reaches its optimal value when the following conditions are met:

$$\begin{cases} \frac{\partial L}{\partial w} = 0 \rightarrow w = \sum_{i=1}^{N} \alpha_i y_i \phi(x_i) \\ \frac{\partial L}{\partial b} = 0 \rightarrow \sum_{i=1}^{N} \alpha_i y_i = 0 \\ \frac{\partial L_1}{\partial \xi_i} = 0 \rightarrow c - \alpha_i - \beta_i = 0, i = 1..N \end{cases} \tag{4}$$

Substitute (4) for (3) we get the following quadratic programming problem

$$\max_{\alpha} Q(\alpha) = -\frac{1}{2} \sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j K(x_i, x_j) + \sum_{i=1}^{N} \alpha_i \tag{5}$$

where $K(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle$ is called the kernel function. Solving this quadratic programming (QP) problem subject to constrains in (4), a decision hyperplane in the high dimensional space can be obtained and will be used in the subsequent classifications.

2.3 The Hybrid Algorithm

The basic idea of the hybrid algorithm is to keep classification branches of a resulting decision tree that have high classification accuracy (corresponds to a significant decision rule) while combining samples that are classified under branches with low classification accuracy into a new training dataset to use the SVM classifier. The modified decision tree classifier is responsible for constructing significant and compact decision rules for human interpretation and the SVM classifier is responsible for training the samples that do not fit in the decision rules of the resulting decision tree. By giving the ill-classified samples in the decision tree classifier a new chance in the SVM classifier, we expect the overall classification accuracy to be higher than using the decision tree classifier alone. The heuristics behind the expectation are as follows. In the decision tree classifier, there are samples in a multi-class training data set, although their patterns may be well perceived by human, they are small in sizes and are often assigned to various branches during the classification processes according to information entropy gain or gain ratio criteria. At some particular classification levels, the numbers of the samples may be below predefined thresholds in decision tree branches to be qualified as decision tree leaf nodes with high classification accuracies, thus the splitting processes stop and they are treated as noises. However, if we combine these samples into a new dataset and train a SVM classifier, since the distribution of the new dataset may be significantly different from the original one, new meaningful patterns may be discovered by the SVM classifier. The basic idea of the hybrid algorithm is illustrated in Fig. 1.

We next present the hybrid algorithm as a set of linked procedures. The overall control flow of the hybrid algorithm is shown in Fig 2. The process of building the modified decision tree classifier is shown in Fig. 3. The process of classifying a sample by the hybrid algorithm is shown in Fig. 4. Since we use a regular SVM classifier, the procedures for building a SVM classifier (Build_SVM) and classifying a sample using the SVM classifier (SVM_Classify) are omitted.

The function *Build_Tree* (Fig. 3) recursively partitions a data set into two and builds a decision tree by finding a partition attribute and its partition value based on the information gain and the gain ratio criteria as discussed previously. There are several parameters used in function *Build_Tree. Min_obj1* specifies the number of samples to determine whether the branches of a decision tree should stop or continue partitioning. *min_obj2* specifies the minimum number of samples for a branch to be qualified as having high classification accuracy. *Min_accuracy* specifies the percentage of samples of the dominating classes. While the purposes of setting *min_obj1* and *min_accuracy* are clear, the purpose of setting *min_obj2* is to prevent generating small branches with high classification accuracies in hope that the samples that fall within the branches can be used to generate more meaningful patterns in the subsequent SVM classifier.

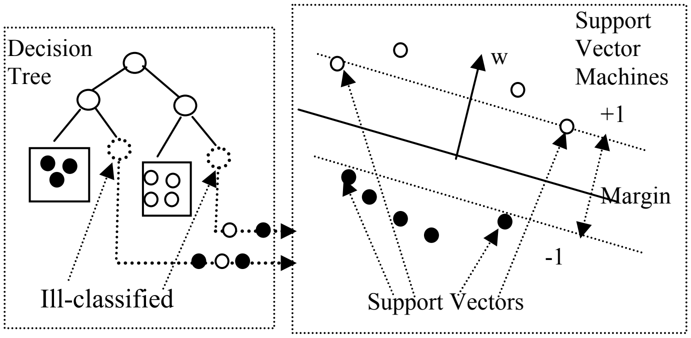


Fig. 1 Illustration of the Basic Idea of the Hybrid Algorithm

Algorithm Hybrid (P, *min_obj1, min_obj2, min_accuracy*)

Inputs:

(1)A training sample dataset (*P*) with *N* samples, each sample has *M* attributes (number of bands of the multi-date images to classify) and a class label

(2)Three thresholds for the modified decision tree algorithm: the number of samples to determine whether the branches of a DT should stop or continue partitioning (*min_obj1*), the minimum number of samples in a branch (*min_obj2*), and the percentage of the samples of classes in branches that can be considered as dominating (*min_accuracy*)

Output:

A hybrid classifier linking a decision tree classifier and a SVM classifier

```
1. Set dataset $D=P$, dataset $D'=\{\}$
2. Call H.T=Build_Tree ( D, D',min_obj1, min_obj2, min_accuracy)
3. Call H.V=Build_SVM(D')
4. Return H
```

Fig. 2 Overall Control Flow of the Hybrid Algorithm

```
Procedure Build_Tree (D, D', min_obj1, min_obj2, min_accuracy)
Inputs:
D': new data set combining ill-classified samples
D, min_obj1, min_obj2, min_accuracy: same as in function Hybrid in Fig. 2
Output: The modified decision tree
 1. Let num_corr be the number of samples of the dominating class in D
 2. if(|D|< min_obj1)
        a. If (num_corr>|D|* min_accuracy) and |D|> min_obj2)
            i. Mark this branch as high accuracy branch (no need for further partitioning) and assign the label of the dominating class
               to the branch
            ii. Return NULL
        b. Else
            i. Mark this branch as low accuracy branch with "use_svm"
            ii. Merge D into D'
            iii. Return NULL
 3. else
        a. if (num_corr>|D|* min_accuracy)
            i. Mark this branch as high accuracy branch (no need for further partition) and assign the label of the dominating class to
               the branch
            ii. Return NULL
//begin binary partition
 4. For each of the attributes of D, find partition value using entropy_gain or gain_ratio
 5. Find the partition attribute and its partition value that has largest entropy_gain or gain_ratio
 6. Divide D into two partitions according to the partition value of the attribute, D1 and D2
 7. Allocate the tree structure to T
 8. T.left_child= Build_Tree(D1, D', min_obj1, min_obj2, min_accuracy)
 9. T.right_child= Build_Tree(D2, D', min_obj1, min_obj2, min_accuracy)
 10. return T
```

Fig. 3 Procedure Build_Tree

If the *min_accuracy* value is set to a high percentage, many branches in the corresponding decision trees will not be able to be qualified as having high classification accuracy and samples that fall within these branches will need to be fed to the subsequent SVM classifier. On the other hand, using higher *min_accuracy* values generates decision branches that are higher in classification accuracies but smaller in numbers. For *min_obj1* and *min_obj2,* it is clear that *min_obj1* needs to be greater than *min_obj2.* The larger *min_obj1*, the earlier to check whether to further partition a decision tree branch. Once the number of samples is below *min_obj1*, the branch will be either marked as having high classification accuracy or marked as needing to be linked to the subsequent SVM classifier, depending on *min_accuracy* and *min_obj2.* A larger *min_obj1,* together with a higher *min_accuracy* makes the hybrid algorithm find larger but fewer decision branches that are high in classification accuracy (i.e., significant decision rules). The parameter *min_obj2* is more related to determining the granularity of "noises" of the decision tree. A smaller *min_obj2* means that fewer branches, the samples of which are almost of the same class (>*min_accuracy)* but are small in sizes, will be considered as unclassifiable in the decision tree classifier and need to be sent to the SVM classifier.

3. Software Implementation

We implement HC-DT/SVM on top of two Java open source data mining packages. The WEKA (Witten and Frank 1999) is a well-known general purpose data mining tool and has been successfully used in GESCONDA - an intelligent data analysis system for knowledge discovery and management in environmental databases (Gibert et al 2006). We use WEKA to provide input/output formatting and use its J48 implementation of the C4.5 algorithm as a skeleton for implementing the DT part of the HC-DT/SVM algorithm. LibSVM (Chang and Lin 2001) is a popular Java library for building SVM classifiers and a wrapper called WLSVM (WEKA LibSVM) has been provided to interface between LibSVM and WEKA (El-Manzalawy and Honavar 2005).

While these open source packages provide building blocks to implement HC-DT/SVM, the hybridization of the two algorithms and providing an integrated implementation is non-trivial for three reasons. First, the J48 code in the WEKA needs to be significantly revised to make it be aware of ill-classified branches. Second, a controlling mechanism needs to be implemented to gather training samples in the ill-classified branches and send them to a SVM classifier. Finally, a new classifier needs to be implemented to dispatch a testing sample to either the modified DT classifier or the SVM classifier and output the combined classification result.

We follow the structure of the weka.classifiers.trees.j48 package and modify the relevant components to implement HC-DT/SVM. First, in addition to *NoSplit* class that represents the leaf node in a constructed decision tree, the *NextSplit* class represents the ill-classified tree branches is added, both extend the *ClassifierSplitModel* class in the J48 package. The *C45ModelSelection* module is extended to handle the new category of decision tree nodes. The instances of the *NextSplit* class always return -1 when a sample is being classified and thus training samples fall under the ill-classified decision tree branches can be gathered and sent to the linked SVM classifier. Similarly, the DT part of the hybrid classification algorithm returns -1 to indicate that a testing sample falls under an ill-classified decision tree branch and should use the linked SVM for final classification. Finally the hybrid classifier implements the *buildClassifier*, *classifyInstance* and *distributionForInstance* interface functions required by a WEKA classifier so that it can be used the same as other WEKA classifiers. By adding a simple component (*NextSplit)*, and slightly revising two existing components (*ClassifierSplitModel, C45ModelSelection),* our implementation of HC-DT/SVM maintains high compatibility with the J48 implementation of the C4.5 decision tree algorithm, which is desirable with respect to minimizing development cost, easy understanding and better usability.

```
Procedure Hybrid_Classify(H, I)
Input:
•   A Hybrid Classifier
•   A sample I with M attributes
Output:
•   Class label of I
 1.   Set label= DT_classify(H.DT,I)
 2.   If label==”use_svm”  return SVM_Classify(H.SVM,I)
 3.   Else Return label

Algorithm DT_Classify (T, I)
Input:
•   A decision tree resulting from the modified decision tree
    classifier
•   An instance I with M attributes
Output:
•   Class label of I
 1.   If T is a leaf node
        a.  If T is marked as a high classification confidence node
                i.  Assign the class label of T to I
               ii.     Return
       b.         Else if T is marked as a low classification
                  confidence node
                     Return “use_svm”
 2.   Else
        a.  Let A be the partitioning attribute and V be the
                partition value
        b.  If(I[A]<=V) then
               Return DT_Classify (T.left_child, I)
        c.  Else
                Return DT_Classify (T.right_child, I)
```

Fig. 4 Procedure *Hybrid_Classify*

We note that, while our implementation of HC-DT/SVM currently takes samples in WEKA's data format only and cannot read data maintained by commercial remote sensing data processing systems, it is possible to use third party open source packages to generate training and testing samples from data managed by the commercial systems and feed the samples to HC-DT/SVM. For example, the StarSpan package developed at the Center for Spatial Technologies and Remote Sensing (CSTARS) at University of California at Davis (Rueda et al 2005). Given a set of images and Regions of Interests (ROIs), StarSpan can extract values and its label of pixels fall within the ROIs and exported them in a variety of data formats which can be further converted to the WEKA's ARFF format.

4. Experiments

To validate the proposed hybrid algorithm, we use a dataset consists of two Landsat TM scenes of southern China acquired on 10 December 1988 (T1) and 03 March 1996 (T2). Preprocessing including geometric and atmospheric corrections of the dataset has been described elsewhere (Seto and Liu 2003). A total of 12 bands, i.e., TM bands 1-5 and band 7 for the two scenes, are used in the classification. Class labels and the numbers of training and testing samples for the classes are listed in Table 1. The six bands in the T1 image are numbered b0 through b5 and the six bands in the T2 image are numbered b7 through b11, respectively.

4.1 Tests of Accuracies

We use the following parameters in the hybrid classifier: min_obj1=200, min_obj2=100 and min_accuracy=95%. For the SVM parameters used in the hybrid classifier, we use a Radial Base Function (RBF) kernel and set G=1 and C=39 after fine tuning. The default parameters in the J48 decision tree implementation are used without fine tuning. For fair comparison, we use the same fine-tuned SVM parameters in the original SVM classifier for the hybrid classifier. The overall accuracy of the hybrid classifier is 89.87%. The classification accuracies for the original decision tree classifier and the SVM classifier are 81.25% and 90.31%, respectively. The error matrices for the three classifiers are listed in Table 2, Table 3 and Table 4, respectively. From the results we can see that the hybrid classifier has much higher accuracies than the classic DT classifier while slightly worse than the SVM classifier.

Table 1 Classes and the numbers of their training and testing samples

Class ID	Class Description	# of Training Samples	# of Testing Samples
1	Water	250	59
2	Natural vegetation	568	154
3	Agriculture	962	246
4	Urban	682	154
5	Water to Urban	544	84
6	Agriculture to Urban	1059	180
7	Vegetation to Urban	775	259
Total		4840	1136

Table 2 Error Matrix of the Hybrid Classifier (Overall Accuracy= 89.88%, Kappa= 0.8784)

	Reference Categories							Σ	UA
	1	2	3	4	5	6	7		
Water	58	0	1	0	0	0	0	59	98.31%
Natural vegetation	0	148	6	0	0	0	0	154	96.10%
Agriculture	0	31	199	1	0	12	3	246	80.89%
Urban	0	0	7	139	0	8	0	154	90.26%
Water to Urban	0	0	0	0	84	0	0	84	100.00%
Agriculture to Urban	0	0	12	3	1	154	10	180	85.56%
Vegetation to Urban	0	0	3	0	0	17	239	259	92.28%
Total								1136	

Table 3 Error Matrix of the Classic DT Classifier (Overall Accuracy=81.25%, Kappa=0.7750)

	Reference Categories							Σ	UA
	1	2	3	4	5	6	7		
Water	46	0	13	0	0	0	0	59	77.97%
Natural vegetation	0	146	8	0	0	0	0	154	94.81%
Agriculture	1	26	188	9	0	15	7	246	76.42%
Urban	0	0	6	134	0	8	6	154	87.01%
Water to Urban	0	0	0	0	84	0	0	84	100.00%
Agriculture to Urban	0	0	9	9	2	136	24	180	75.56%
Vegetation to Urban	0	4	8	0	0	58	189	259	72.97%
Total								1136	

Table 4 Error Matrix of the Classic SVM Classifier (Overall Accuracy =90.32%, Kappa=0.8838)

	Reference Categories							Total	UA
	1	2	3	4	5	6	7		
Water	58	0	1	0	0	0	0	59	98.31%
Natural vegetation	1	149	4	0	0	0	0	154	96.75%
Agriculture	0	27	202	0	0	10	7	246	82.11%
Urban	0	0	4	145	0	5	0	154	94.16%
Water to Urban	0	0	0	0	84	0	0	84	100.00%
Agriculture to Urban	0	0	12	7	1	153	7	180	85.00%
Vegetation to Urban	0	0	2	0	0	22	235	259	90.73%
Total								1136	

Table 5 Accuracy Comparisons of the Hybrid, the Classic DT and the SVM Classifiers

	Accuracies			Hybrid~DT Test		Hybrid~SVM Test	
	Hybrid	DT	SVM	Z-Score	Significance Level	Z-Score	Significance Level
Water	98.31%	77.97%	98.31%	3.4162	P<0.001	/	/
Natural vegetation	96.10%	94.81%	96.75%	0.5471		-0.307	
Agriculture	80.89%	76.42%	82.11%	1.2104		-0.3483	
Urban	90.26%	87.01%	94.16%	0.8977		-1.2754	
Water to Urban	100.00%	100.00%	100.00%	/	/	/	/
Agriculture to Urban	85.56%	75.56%	85.00%	2.397	P<0.01	0.1487	
Vegetation to Urban	92.28%	72.97%	90.73%	5.7982	P<0.001	0.6304	
Overall	89.88%	81.25%	90.32%	5.8499	P<0.001	-0.3512	

To further compare the classification accuracies at the individual class level, the accuracies for the hybrid, classic DT and SVM classifiers for each of the seven classes are listed in Tables 2-4 as well. Z-statistics between the accuracies of the hybrid classifier and the classic DT classifier and Z-statistics between the accuracies of the hybrid classifier and the SVM classifier for the classes are also calculated and listed in Table 5. For classifications using two classifiers and having the same accuracies, it is not possible to calculate Z-statistics and the correspondingly Z-scores and confidence levels are marked with "/". From the results it is clear that the hybrid classifier outperforms the classic decision tree classifier for all classes except *Water to Urban* where both classifiers achieve full (100%) classification accuracies. More specifically, the hybrid classifier outperforms the classic DT classifier for classes *Agriculture to Urban* at p<0.01 significance level and *Water* and *Vegetation to Urban* at p<0.001 significance level. Table 5 also shows that while the SVM classifier achieves slightly better with respect to the overall classification accuracy than the hybrid classifier, SVM is not always better than the hybrid classifier at the class level. In fact, the SVM classifier performs better only for four out of the seven classes and none of them are statistically significant at the significance level p<0.1.

4.2 Test of Interpretability

While the hybrid classifier achieves much higher classification accuracies than the classic DT classifier and comparable classification accuracies to the SVM classifier, the most significant advantage of the hybrid classifier is its capability to generate concise and human interpretable decision rules. Among the 4840 training samples, the hybrid classifier generalizes 2141 samples and creates a compact decision tree (Fig. 5). The resulting decision tree has eight leaves which can be easily translated into decision rules. In contrast, the decision tree resulting from the original decision classifier has 214 leaves and is too big to fit in a page for presentation. In addition, we find that the significant decision rules resulting from the classic DT classifier are mixed with insignificant decision rules and it is hard for users to interpret. Thus the hybrid classifier has the capacity to leverage the most significant decision rules with high classification accuracies and present them to users for immediate validations.

```
b4 <= 10
|  b11 <= 8
|  |  b9 <= 12: Water (165.0)
|  b11 > 8
|  |  b3 <= 17: Water to Urban (387.0/6.0)
b4 > 10
|  b8 <= 40
|  |  b0 <= 66
|  |  |  b7 <= 29: Natural vegetation (342.0/9.0)
|  |  b0 > 66
|  |  |  b6 > 68
|  |  |  |  b5 <= 17: Agriculture (387.0/14.0)
|  |  |  |  b5 > 17
|  |  |  |  |  b9 > 38: Agriculture (222.0/9.0)
|  b8 > 40
|  |  b2 <= 51
|  |  |  b3 > 36
|  |  |  |  b7 > 38
|  |  |  |  |  b1 <= 30
|  |  |  |  |  |  b10 > 111
|  |  |  |  |  |  |  b1 <= 28
|  |  |  |  |  |  |  |  b5 <= 18: Vegetation to Urban (185.0/5.0)
|  |  |  |  |  b1 > 30
|  |  |  |  |  |  b3 <= 43
|  |  |  |  |  |  |  b3 <= 40: Agriculture to Urban (116.0/1.0)
|  |  b2 > 51: Urban (337.0/5.0)
```

Fig. 5 The Resulting Compact Decision Tree

The resulting compact decision tree from the hybrid classifier is quite meaningful. The first decision rule, b4<=10 and b11<=8 and b9<=12➔*Water* generalizes 165 out of the 250 training samples of the class (66.0%). The second rule, b4<=10 and b11>8 and b3<=17 ➔ *Water to Urban* generalizes 387 out of the 544 training samples (69.3%) with six exceptions. The exceptions are allowed because the min_accuracy is set to 95% and there could be up to 5% exceptions. By comparing the two rules, it is easy to derive the following interpretations. Class *Water* has low values in both band 5 at time T1 image (b4) and band 7 at time T2 image (b11). While class *Water to Urban* has low values in band 5 at time T1 image (b4), it has high values in band 7 at time T2 image (b11). The derived rules match domain knowledge very well - urban samples (pixels) have higher values than water samples. This can be further explained by the rule derived from the very bottom branch of the decision tree in Fig. 5 related to class *Urban*: b4>10 and b8>40 and b2>51➔*Urban*. The rule generalizes 337 out of the 682 samples (49.4%) with just five exceptions.

Similarly, the following five rules can be derived for the rest four classes:

1) B4>10 and b8<=40 and b0<=66 and b7<=29➔*Natural Vegetation*. The rule generalizes 342 out of the 568 samples (60.2%) with 9 exceptions.

2) B4>10 and b8<=40 and b0>66 and b6>68 and b5<=17➔*Agriculture*. The rule generalizes 342 out of the 962 samples (40.2%) with 14 exceptions

3) B4>10 and b8<=40 and b0>66 and b6>68 and b5>17 and b9>38➔*Agriculture*. The rule generalizes 222 out of the 962 samples (23.1%) with 9 exceptions

4) B4>10 and b8>40 and b2<=51 and b3>36 and b7>38 and b1<=30 and b1<=28 and b5<=18➔ *Vegetation to Urban*. The rule generalizes 185 out of the 775 samples (23.9%) with 5 exceptions.

5) B4>10 and b8>40 and b2<=51 and b3>36 and b7>38 and b1>30 and b3<=40➔ *Agriculture to Urban*. The rule generalizes 116 out of the 1059 samples (11.0%) with 1 exception.

Rule 2 and rule 3 are related to the same class (*Agriculture*). If we group the two rules then 564 out of the 962 samples (58.6%) can be generalized with 23 exceptions. The two rules for the *Agriculture* class are pretty similar and fall in the same decision tree branch (B4>10 and b8<=40 and b0>66). The breaching point for class *Agriculture* and class *Natural Vegetation* is b0=66 which indicates that natural vegetation has lower pixel values than agriculture at band 1 in time T1 image (b0). Compared with the rules of the other five classes, rules representing the two change classes *Vegetation to Urban* and *Agriculture to Urban* (Rule 4 and Rule 5) are less well represented since lower percentages of the samples of the classes can be generalized by the rules. This might indicate that these two classes are more complex and their sample values may not fit linear classifiers (such as decision tree) very well. Similar to characterizing the differences between class *Water* and class *Water to Urban* (and class *Natural Vegetation* and class *Agriculture* as well), from the resulting decision tree it is clear that, the difference between the samples that are generalized by Rule 4 (*Vegetation to Urban*) and Rule 5 (*Agriculture to Urban*) is that *Vegetation to Urban* has smaller values at band 2 of time 1 image (b1) than these of class *Agriculture to Urban*. The breaching point is b1=30. However, since the samples generalized by the two decision rules are only a fraction of the total samples of the two classes, cautions are needed to validate this interpretation.

The resulting decision tree also naturally generates a hierarchy of the seven classes in the change detection dataset. From Fig. 5 we can see that water related classes (*Water* and *Water to Urban*) are first separated from the rest. The clustering process is followed by grouping urban related classes (*Vegetation to Urban*, *Agriculture to Urban* and *Urban*) as one cluster and *Natural Vegetation*/*Agriculture* as another cluster. Among the cluster for urban related classes, the two changing classes (*Vegetation to Urban* and *Agriculture to Urban*) are naturally grouped again. The class hierarchy can be used for knowledge transfer (Rajan and Ghosh 2006) and to refine the classification process. For example, decomposing a multi-class problem into multiple binary classifications based on the class hierarchy.

5. Conclusions and Future Work

In this study, we have proposed a hybrid algorithm that tightly integrates a decision tree algorithm and a SVM algorithm to classify multi-date images for land cover change detections. Experimental results show that the hybrid algorithm significantly improves the accuracies of the classic decision tree based classifier and achieves comparable classification accuracy to classic SVM based classifier. In addition, the hybrid algorithm leverages the most significant decision rules with high classification confidences and presents them to user for immediate evaluations.

The proposed hybrid algorithm represents a framework of hybridizing existing classification algorithms for classifying remotely sensed images. Due to the fuzzy and vague nature of classes defined by human and the inaccuracy introduced by the sampling process, the existence of samples that are difficult to classify is inevitable. Instead of using a single complex classifier for all the samples, it is more beneficial to use simple classifiers for "easy" samples and generate human interpretable knowledge from the classifiers through visualization (Zhang et al 2009) while leave the "difficult" samples for more sophisticated classifiers where visualization is usually not available. We also would like to point out that, while this work originates from the multi-date land cover change detection research, HC-DT/SVM is generic enough to be applied to a variety of types of environmental data analysis where supervised classifications are involved.

For future work, first, we would like to incorporate the hybrid classification algorithm into our VDM-RS (Visual Data Mining for Remote Sensing) prototype system (Zhang et al 2009) and help users gain more insights into the data, classification algorithm and results through visualization, interaction and exploration. Second, we would like to compare the HC-DT/SVM algorithm with other approaches discussed in the introduction section and test them on additional datasets. Finally, we plan to release the implementation as an open source package after proper documentation.

6. Acknowledgements

The author thanks Dr. Weiguo Liu for providing the test dataset and constructive discussions.

7. References

1. Chan, J. C. W., Chan, K. P. and Yeh, A. G. O., 2001. Detecting the nature of change in an urban environment: A comparison of machine learning algorithms. Photogrammetric Engineering and Remote Sensing 67(2): 213-225.
2. Chang, C. and Lin, C. 2001. LIBSVM: a Library for Support Vector Machines, http://www.csie.ntu.edu.tw/~cjlin/libsvm.
3. De Colstoun, E. C. B. and Walthall, C. L., 2006. Improving global scale land cover classifications with multi-directional POLDER data and a decision tree classifier. Remote Sensing of Environment 100(4): 474-485.
4. De Fries, R. S. and Chan, J. C. W., 2000. Multiple criteria for evaluating machine learning algorithms for land cover classification from satellite data. Remote Sensing of Environment 74(3): 503-515.
5. EL-Manzalawy, Y. and Honavar, V. 2005. WLSVM: Integrating LibSVM into Weka Environment, http://www.cs.iastate.edu/~yasser/wlsvm.
6. Friedl, M. A. and Brodley, C. E., 1997. Decision tree classification of land cover from remotely sensed data. Remote Sensing of Environment 61(3): 399-409.
7. Friedl, M. A., Brodley, C. E. and Strahler, A. H., 1999. Maximizing land cover classification accuracies produced by decision trees at continental to global scales. IEEE Transactions On Geoscience and Remote Sensing 37(2): 969-977.
8. Friedl, M. A., McIver, D. K., Hodges, J. C. F., Zhang, X. Y., Muchoney, D., Strahler, A. H., Woodcock, C. E., Gopal, S., Schneider, A., Cooper, A., Baccini, A., Gao, F. and Schaaf, C., 2002. Global land cover mapping from MODIS: algorithms and early results. Remote Sensing of Environment 83(1-2): 287-302.
9. Gibert, K., Sanchez-Marre, M. and Rodriguez-Roda, I., 2006. GESCONDA: An intelligent data analysis system for knowledge discovery and management in environmental databases. Environmental Modelling & Software 21(1): 115-120.
10. Huang, C., Davis, L. S. and Townshend, J. R. G., 2002. An assessment of support vector machines for land cover classification. International Journal of Remote Sensing 23(4): 725-749.
11. Huang, Z. and Lees, B. G., 2004. Combining non-parametric models for multisource predictive forest mapping. Photogrammetric Engineering and Remote Sensing 70(4): 415-425.
12. Im, J. and Jensen, J. R., 2005. A change detection model based on neighborhood correlation image analysis and decision tree classification. Remote Sensing of Environment 99(3): 326-340.
13. Kelly, M., Shaari, D., Guo, Q. H. and Liu, D. S., 2004. A comparison of standard and hybrid classifier methods for mapping hardwood mortality in areas affected by "sudden oak death". Photogrammetric Engineering and Remote Sensing 70(11): 1229-1239.
14. Liu, W. G., Gopal, S. and Woodcock, C. E., 2004. Uncertainty and confidence in land cover classification using a hybird classifier approach. Photogrammetric Engineering and Remote Sensing 70(8): 963-971.

15. Lu, D., Mausel, P., Brondizio, E. and Moran, E., 2004. Change detection techniques. International Journal of Remote Sensing 25(12): 2365-2407.
16. Lu, D. and Weng, Q., 2007. A survey of image classification methods and techniques for improving classification performance. International Journal of Remote Sensing 28(5): 823-870.
17. Nemmour, H. and Chibani, Y., 2006. Multiple support vector machines for land cover change detection: An application for mapping urban extensions. Isprs Journal of Photogrammetry and Remote Sensing 61(2): 125-133.
18. Pal, M. and Mather, P. M., 2005. Support vector machines for classification in remote sensing. International Journal of Remote Sensing 26(5): 1007-1011.
19. Pal, M. and Mather, P. M., 2006. Some issues in the classification of DAIS hyperspectral data. International Journal of Remote Sensing 27(14): 2895-2916.
20. Prasad, A. M., Iverson, L. R. and Liaw, A., 2006. Newer classification and regression tree techniques: Bagging and random forests for ecological prediction. Ecosystems 9(2): 181-199.
21. Rajan, S., Ghosh, J. and Crawford, M. M., 2006. Exploiting class hierarchies for knowledge transfer in hyperspectral data. IEEE Transactions on Geoscience and Remote Sensing 44(11): 3408-3417.
22. Rueda, C. A., Greenberg, J. A. and Ustin, S. L., 2005. StarSpan: A Tool for Fast Selective Pixel Extraction from Remotely Sensed Data. Center for Spatial Technologies and Remote Sensing (CSTARS), University of California at Davis, Davis, CA.
23. Seto, K. C. and Liu, W. G., 2003. Comparing ARTMAP neural network with the maximum-likelihood classifier for detecting urban change. Photogrammetric Engineering and Remote Sensing 69(9): 981-990.
24. Singh, A., 1989. Digital Change Detection Techniques Using Remotely-Sensed Data. International Journal of Remote Sensing 10(6): 989-1003.
25. Witten, I. H. and Frank, E., 2000. Data Mining: Practical machine learning tools with Java implementations. San Francisco, CA, Morgan Kaufmann.
26. Wynne, R. H., Joseph, K. A., Browder, J. O. and Summers, P. M., 2007. Comparing farmer-based and satellite-derived deforestation estimates in the Amazon basin using a hybrid classifier. International Journal of Remote Sensing 28(6): 1299-1315.
27. Yuan, F., Sawaya, K. E., Loeffelholz, B. C. and Bauer, M. E., 2005. Land cover classification and change analysis of the Twin Cities (Minnesota) Metropolitan Area by multitemporal Landsat remote sensing. Remote Sensing of Environment 98(2-3): 317-328.
28. Zhang, J., Liu, W. and Gruenwald, L. (2007). A Successive Decision Tree Approach to Mining Remotely Sensed Image Data. Knowledge Discovery and Data Mining: Challenges and Realities. X. Zhu and I. Davidson, Idea Group Publishing, Inc: 98-112.
29. Zhang J.,Gruenwald, L. and Gertz. M (2009). VDM-RS: A Visual Data Mining System for Exploring and Classifying Remotely Sensed Images. Computers & Geosciences 35(9), 1827-1836

Framework of Integration for Collaboration and Spatial Data Mining Among Heterogeneous Sources in the Web

André Fabiano de Moraes
Federal Catarinense Institute , Department of Information Technology
PO Box 16, 88340-000
Camboriú, SC, Brazil
Phone (+55 47 2104-0800)
ecv3afm@ecv.ufsc.br

Lia Caetano Bastos
Federal University of Santa Catarina, Department of Civil Engineering
PO Box 476, 88040-970
Florianópolis, SC, Brazil
Phone (+55 47 3721-7099)
ecv1lcb@ecv.ufsc.br

ABSTRACT

This paper highlights the diversity of spatial data of rural and urban properties, constantly generated by different public institutions, as well as the existing problems of exchange of information among them. Firstly, this work describes the results obtained in the study and development of an agile flexible method to offer support construction, implementation and accompaniment activities of free geo-solutions for the web, aiming at a growing community of users and developers who manipulate geographic data. Next, the development of the OpenICGFw (Integration for Collaborative Geospatial Framework for the Web) that seeks, through a single environment to assist in the integration and collaboration among different sources of spatial data in synchrony with the efforts and specifications of OGC and W3C. To do this, the evaluation study for the construction of the framework is presented where it was possible to apply MCDA-C (Multi Criteria Decision Aiding – Constructivist) in the identification of the fundamental and elementary aspects for the construction of the framework. Details are presented by means of a case study that illustrates data exported from different geospatial information systems requiring the integration of census, environmental, urban and rural information over the internet. During the discussion the results obtained using this framework are presented, providing, through web mapping applications, the implementation of collaborative strategies seeking the integration of bases distributed for the use of spatial data mining techniques.

Categories and Subject Descriptors

D.2.6 [Software]: Programming Environments - *Integrated environments*; H.2.8 [Database Management]: Database Applications - *data mining, scientific databases, spatial databases and GIS*; H.3.5 [Information Systems]: Online Information Services: *data sharing*.

General Terms

Interoperability, Collaboration and Spatial Data Mining.

ACM SIGSPATIAL DMG Workshop, Nov 2, San Jose, CA, U.S.A.
Copyright 2010 ACM 978-1-4503-0430-6/10/11...$10.00.

Keywords

Framework; Collaboration; Spatial Data Mining; Development of Free Software.

1. INTRODUCTION

With the expansion of computer networks and the growing use of geo-processing technologies as a tool for decision making in several areas, a strong involvement has been seen from public institutions in the processes of spatial analysis. However, the lack of tools which facilitate integration and interoperability between different sources of geographic data has hindered these processes, along with the infrequent use and implementation of existing software due to lack of resources that normally affect public and private sectors [1]. Associated with this, is the access to geographic data via the Internet and all problems relating to its exchange due to the particular nature of each set of data and their definition.

To this end, some challenges are cited in [2], such as the complexity and lack of data standards, such as the interoperability between spatial data, and especially the lack of common conceptual models presenting problems in data exchange between distinct geographic information systems (GIS). Focusing on this some research has presented experiments to minimize or overcome the problems, as reported in [2], [3] and [4]. In heterogeneous system environments, data conversion represents a cost of between 60% and 80% of the total cost of implantation.

From this point of departure, a method involving agile flexible aspects for assisting elaboration, development, construction, implementation, and also monitoring activities, for free geo-solutions for the web is presented in this paper. For definition of this method concepts of the MCDA-C (Multicriteria Decision Aid - Constructivist), proposed in [5], have been used enabling the assessment of resources available in the main free softwares adopted in geographic information systems, as well as implementation of new features (Section 1.1).

For this, a flexible model of development is necessary in order to address issues related to the constants collected for variable requirements [6], in addition to specifications for data patterns and techniques commonly used.

The collaborative development of open source software is a very attractive alternative [7], on the other hand, it also arouses some uncertainty on the part of institutions. However, it could significantly assist the institutions in the acquisition of technologies such as the elaboration of successful projects by well-organized communities.

The primary motivation for this work arises because of the lack of integration of data between computer systems in public institutions. The latter generating poor information and the resulting clash of interests in providing services to society. Nevertheless, this paper offers two contributions aiming to overcome these deficiencies. Firstly the requirements for preparing the OpenICGFw are assessed using MCDA-C. Subsequently, the OpenICGFw addressing the stages of development (section 3) is presented and its applicability examined as a case study (section 4).

1.1 Structuring Problems for Multicriteria Decision Analysis (MCDA-C)

Initially in [8] the process of decision support is an open system with specific components, such as actors, values, goals, actions and features. The support decision process could then be seen as an interactive process with the problem of poor-structure, where the elements and their relationships emerge somewhat chaotically.

Also in [5] the structuration of the decision support process aims at building a structure acceptable to the actors modeling a preexisting reality. This is the most important phase of MCDA-C, as it provides learning, clarity and representation by defining and constructing a model that will serve as a common base, where the values of the actors involved can be validated. The purpose of structuration is to develop a set of tools that enable decision makers to better understand the problem.

That is to say, the existence of an unsatisfactory factor, a proponent, the importance of which warrants the effort to resolve the problem and finally that it is actually possible to resolve [5]. In order to define the dimensions to be worked on for the construction of a strategic model, each elementary point of view (EPV) is initially introduced; they represent a set of minimum requirements for the structuration and functioning of the method.

From the results obtained from the structuration of the method, in line with the proposals from the Georeferencing Theme Group - Electronic Government [9], within the National Spatial Data Infrastructure (INDE) [10] and also with the specifications of standardization of spatial data proposed by the OGC [11], it was possible to support the development of FVPs (Fundamental Points of View) and (Elementary Points of View). This facilitates the understanding of the requirements and prioritizing the development of a strategic environment for integration and collaboration.

1.2 Defining Standards for the Method of Development

After the structuration of the problems, they are associated to the method of developing specific criteria to guide the implementation of the method. Many of these criteria are drawn from other disciplines, such as object orientation [12 and 13] and web engineering [14]. Other features are simplified and adapted to the process of developing software proposed in [6] in an agile and flexible manner.

From this starting point, construction work on the free software called OpenICGFw (Open Integration Collaborative Geospatial Framework for the Web) began. This has been developed using the MVC standard (Model View Controller) [15, 16, 17, 18, 19, 20 and 21] for operation in the web environment. This facilitates the inclusion of new concepts for the production of webmapping systems, including MVCS (Model, View, Controller and Services) which are also presented in [22].

In [23] a framework aggregates common functionality across multiple applications and makes them available in structures that aim to be easy to handle and understand. For this purpose surveys were also carried out on the development of frameworks commonly used by the free software community, seeking to provide an investigation of points already established, aiding in the analysis and formulation of key aspects for modeling a new environment.

Moreover, such aspects should be thoroughly investigated and pre-selected aiming at the reuse of existing functions and made freely available for use, alteration and adaptation, varying according to the individual needs of each project. Each framework above has many advantages that can be used, and the following topics present the process of developing the strategic model of collaboration as well as the initial steps for constructing a webmapping framework to manipulate and analyze spatial data via the Internet network. Structuration of the Requirements for the Development of the OpenICGFw Framework Collaborative, are divided in two aspects, the Fundamental and the Elementary.

Fundamental point of view: Aspect of Infrastructure (Physical and Logical Structure); Aspect of Interoperability (Exchange, Catalog of Structures, Analyses); Aspect of Collaboration; Aspect Organizational (Administration).

Elementary point of view: Computer (Hardware); Internet, intranet or Extranet Networks; Operational System; Development Language; DBMS – Database Management Systems; Navigation Interface; OGC Standardization; ISO Standardization; External Libraries (GDAL and TerraLib); Data Types Vector (DGN, DXF, DWG, SHP), Raster (TIF, GeoTIFF, JPG, PNG, GIF) Definition (GeoBR, XML/GML); Metadata; Queries Processing; Statistics; Data Mining.

2. ARCHITECTURE CONSTRUCTING THE OPENICGFW

According to [24] the need for decentralization to manage geographic information systems is increasingly driving the community that develops these systems for desktops [25] to make architectures more flexible for distribution on the Internet network [26].

Unlike other types of information services, this requires specific software environments [27] which enable closer contact between institutions through dynamic communication mechanisms.

For better understanding, we present the main frameworks used in the development of free applications for the web which primarily adopt PHP. Among them are cited: CakePHP (MIT), CodeIgniter (GNU-GPL v.2), Drupal (GNU-GPL), Joomla (GNU-GPL), Miolo (CC-GNU GPL), Symfony (MIT), Zend Framework (BSD) and Zoop Framewok (ZPL). Also specific frameworks for the treatment and provision of spatial data, developed in different languages such as geodjango (BSD), Mapstraction (BSD), OpenGTS (Apache v.2) and OpenStreetMap (GNU-GPL).

According to [26 and 28] to classify the operational relations between client, server, geographic objects and operations among GISs requires detailing via metadata, enabling communication between different services delivered.

To do so, adopting a work strategy through a single environment provides collaboration aimed at defining a set of functions required to solve common problems, as well as support for a particular individual positioning.

It can be observed that this technological strategy responds to long-term issues and that the technology is treated very broadly, i.e. involves issues related to (technical and organizational) products and processes. To develop the method nine stages were elaborated, only the first two stages are contextualized in this paper, with the main objective of initially assisting in the planning and development of new free frameworks as shown in Figure 1.

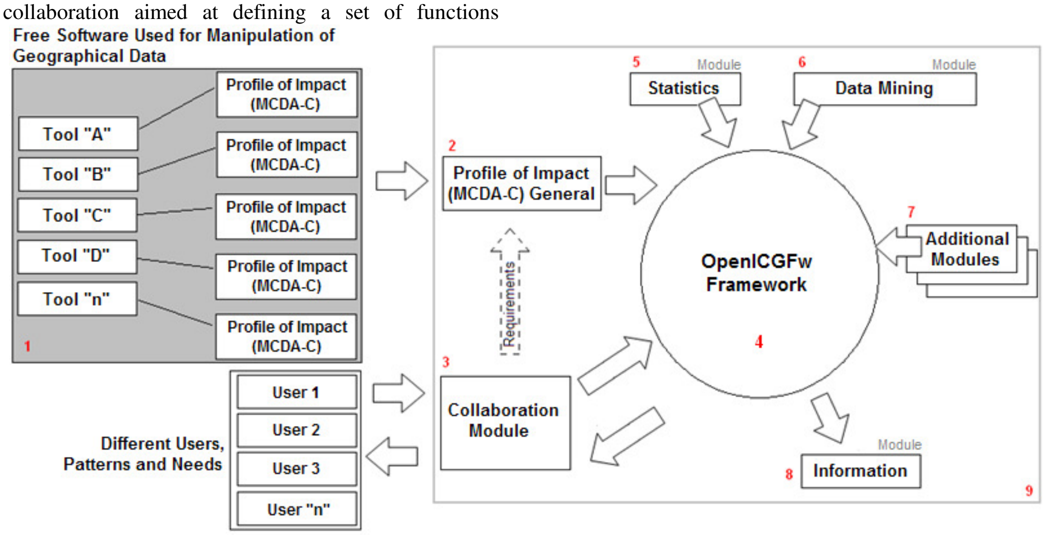

Free Software Used for Manipulation of Geographical Data
Tool "A"
Tool "B"
Tool "C"
Tool "D"
Tool "n"
Profile of Impact (MCDA-C)
Profile of Impact (MCDA-C)
Profile of Impact (MCDA-C)
Profile of Impact (MCDA-C)
Profile of Impact (MCDA-C)
1
Different Users, Patterns and Needs
User 1
User 2
User 3
User "n"
5
Module
Statistics
6
Module
Data Mining
2
Profile of Impact (MCDA-C) General
Requirements
OpenICGFw Framework
4
7
Additional Modules
3
Collaboration Module
Module
8
Information
9

Figure 1: Architecture of the method for construction of the OpenICGFw.

3. DETAILS OF THE PHASES OF DEVELOPMENT OF THE OPENICGFW

This method is divided into three major phases, which first address the key steps including (1st, 2nd, 3rd and 4th) established for the process of elaboration. While the second phase deals with the applications (in the 5th, 6th, 7th and 8th stages). Finally the third phase deals exclusively with the mechanism of assessment potential (Stage 9) of the framework. In the first phase, after the construction of the descriptors with their respective values, a comparison is carried out which aims at assessing the main features and functionalities of the softwares in order to identify the profile of each object of investigation

This provides the generation of a profile of impact. For illustration, four softwares are assessed, chosen by means of user surveys as follows: SpringWeb (http://www.inpe.br), AlovMap (http://alov.org/), CartoWeb (http://cartoweb.org/), I3GEO (http://mapas.mma.gov.br/) and Mapstraction (http://mapstraction.com/).

3.1 Descriptors – First-Stage

The construction of descriptors to specifically aid in the identification of an individualized profile of existing technologies (softwares) was carried out together with users and decision makers aiming at exactly complying with each elementary point. With this it was possible to select 18 descriptors, resulting in: D1-Compatibility with Hardware (Hard); D2-Compatibility with the Internet, Intranet and Extranet Networks (Network); D3-Portability between Operating Systems (OS); D4-Programming Language Adopted (LP); D5-Database Management System (DBMS); D6-Navigation Interface (Nav); D7-Implementing OGC Standard (OGC); D8-Implementing ISO Standard (ISO); D9-Incorporating libraries external to the framework (Bibl); D10-Exchange of geospatial data (Inter); D11-Catalogue of Structures - Metadata (Metad); D12-Process Ontology (Ontol); D13-Enabling integration and the relationship between data (Norm); D14-Providing elaboration and processing of queries (Con); D15-Providing mechanisms for statistical analyses (Stat); D16-Providing Data Mining (DM); D17-Collaboration (Colab) and D18-Organization (Orga).

$$\sum = \frac{((fx_i . a) + (fy_i .b) + (fz_i .c))}{10}$$

fx = Frequency X i = 1,2,…n
fy = Frequency Y i = 1,2,…n
fz = Frequency Z i = 1,2,…n

Equivalence Factor (a) = 5 “Excellency”
Equivalence Factor (b) = 3 “Stable”
Equivalence Factor (c) = 2 “Compromising”

3.2 Assessment of Impact Profile - Second Stage

In the second stage a profile of widespread impact, was noted i.e. the profile that presents a balance between all softwares submitted to the assessment process. The union of essential elements was also obtained, such as the indicators (EVPs), to stipulate the main objectives and priorities in the formulation and development of activities to construct a new framework.

Using Figure 2, a comparison is done between the average profile (m) of the impact of all assessments performed in the first stage, as well as the representation of the best profile with the highest score and also the representation of the status “quo” to be achieved. From these descriptors individual analyses of the architecture and the resources available were carried out for each software.

These analyses enabled the completion of mapping of determining features, enabling the identification of the current profile for each software in relation to levels of Excellency (green), Stability (blue) or Compromising (yellow).

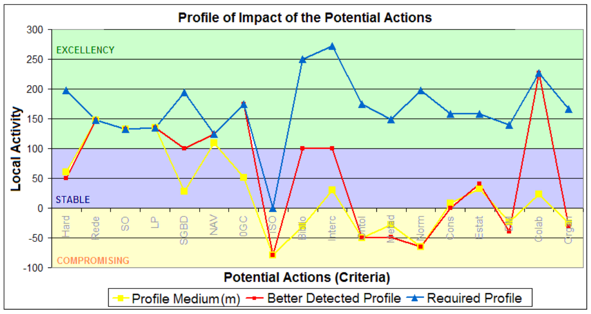


Figure 2: Analysis of each Profile in relation to the Status " Quo " required.

From the intersection of each profile, it was possible to detect which basic points should be prioritized in the execution of activities in the construction of the framework. According to the values extracted at the intersection of each profile we applied the same calculation formula used for ranking the software analyzed in the first stage, and it was possible to visualize the new values and quantify the gains for each improvement, implemented in the construction of the framework. In Table 1 comes the distribution of the respective values for obtaining the final punctuation of each profile.

Table 1: Assessment for the final ranking

Profile	EX	ST	CO	Total	PU **
Profile Medium (m)	4	7	7	18	5,50
Better Detected Profile	9	3	6	18	6,60
Wanted Profile	17	1	0	18	8,80

EX (Excellency), ST (Stability), CO (Compromising) and PU** (score obtained after applying the formula to determine the final ranking).

3.3 Elaboration of the Core of the Framework - Fourth Stage

In this stage the codification of the OpenICGFw (Integration for Collaborative Geospatial Framework for the Web) is performed, taking into account the earlier discussions and the primary requirements obtained for the principal core of the respective framework. The language was chosen for two reasons: the first in compliance with policies, premises and technical specifications defined by Standards of Interoperability for Electronic Government [9] - e-PING, proposed by the Brazilian Government through laws and decrees, the other reason relates to the expressive community of users distributed around the world who use and assist in support for maintaining a stable technology.

For better management of the framework the structuration of a main archive was undertaken featuring a fixed base for storing relational data, enabling the parameterization of functionalities for personalizing the environment. Initially, the archive was also designed aiming at centralizing specific data from processes that are run directly on the framework.

The archive database has been made compatible between two different DBMSs available under the GNU-GPL license, PostgreSQL 8.3, and also for MySQL 5.0. This is necessary due to several features already implemented in both of the DBMSs and principally due to the automated resources for handling geographic data [33]. The possibility of developing applications in three layers (client, server and database), enabled the emergence of various Webmapping systems [34] presenting important components for their use by public institutions.

For documentation of the architecture of the project numerous documentations have been prepared by means of artifacts containing specifications from (Unified Modeling Language) [35] UML 2.0.

3.4 Applied Modularization - Advanced Stages

The modularization has been idealized to provide, through the framework, the integration of basic and also advanced resources. In the fifth stage it contains the development of the statistical module for consultation.

The sixth stage brings the implementation of the data mining module for the extraction of rules in pursuit of knowledge hitherto unobserved. In the seventh stage, the development module assisted in creating new resources through the implementation of scripts. In the eighth stage it was possible to program the module for viewing the results through text, tables and cartograms.

Finally, in the ninth and final stage the module to manage the processes run was implemented with the framework, as well as a panel for its administration.

4. APPLICATION OF THE OPENICGFW – A CASE STUDY

The case study presents the experiments performed to demonstrate the implementation of all processes for handling data from the framework, using spatial data covering mainly the region of the city of Itajaí, in the state of Santa Catarina - Brazil as illustrated in Figure 3.

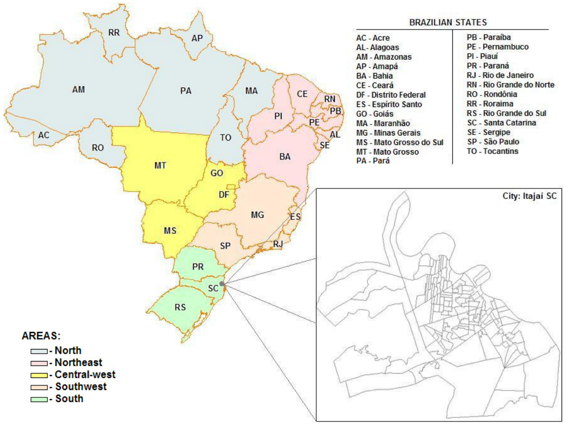


Figure 3: Identification of the area and territorial dimension.

The data has been obtained from distinct sources and is periodically subject to updates of information, the IBGE (Brazilian Institute of Geography and Statistics) [36] with the census data freely available in ".SHP" format, the Local Government of Itajaí [37] has data obtained from the CTU (Urban Technical Registration), offering them partially, Epagri (Agricultural Research and Rural Extension of Santa Catarina) [38] with partial data on mapping of rural areas, for example topography, hydrography, elevation, and others.

Through the INPE (National Institute for Space Research) [39] with satellite images collected for monitoring natural phenomena and their temporality, and the MMA (Environment Ministry) [40] and environmental data.

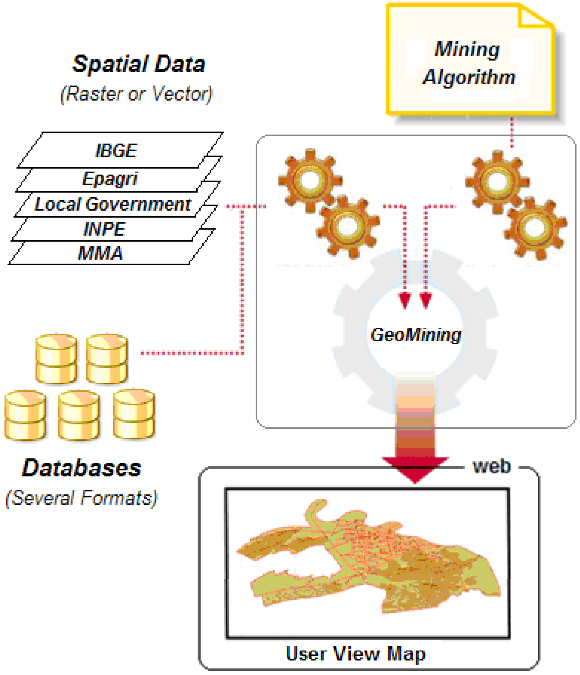


Figure 4: Proposed Architecture for computational processing.

With all this data the objective is the implementation of the framework to discover knowledge not previously addressed by the institutions regarding the overlap of this information as shown in Figure 4. After identifying the area to be used for implementing the case study, it was necessary to get every detail about the source of the collected data, or metadata, as it is extremely important for proper processing. After the registration of institutions and also the proper cataloging of spatial and non spatial data for the sequence of the experiment it was necessary to transform some data into a specific notation, enabling its suitability and standardization for interpretation using the framework.

The next sections bring details of the activities that address the bringing together of geographic data with interoperability, addressing topics that seek to reduce the gap between the projects developed in distinct standards.

4.1 Promoting Interoperability Between Geographical Data

To promote the integration of data through the OpenICGFw some situations are initially restricted, this article specifies a simple way to integrate information as in [41 and 42] to generate updated knowledge for collaboration, planning and coordination of actions among public and also private institutions.

But, at present some challenges are still being faced, such as data stored in workstations and not on a specific server; essential data and products non-catalogued; also the lack of physical and technological infrastructure limiting the storage of large amounts of data.

This requires cataloging and constant updating that promotes the identification of all the body of local and distributed data, producing metadata for digital archives. The experiment with the implementation of GeoNetwork is cited in [43], as a free solution and open source for managing this data, combined with new parameters and functionalities. This enables handling of this data for knowledge discovery by means of cataloging data, such as the monitoring of territorial divisions, environmental data, social inclusion and also regional sustainable development.

An architecture composed of several procedures has been adopted that makes this possible from its authorization to the transformation of specific data for its final visualization. After authorization, for the transfer to the server (optional), and identifying the processing of input data, metadata is needed to better specify the features of one or more sets of data in addition to its storage, either local or remote.

From that diagnosis it is possible to proceed with the running of the other procedures in addition to new modules as shown in Figure 5.

One of the important differentials of this work is the module for the exchange of geographic data, which aims to provide access to these resources through a web interface in a simplified form for the user.

For this, knowing in advance the source of the data, by means of its metadata, facilitates the identification and exchange of previously known structures, significantly reducing the loss of any original features.

In particular some ETL (Extract, Transform, Load) solutions for the exchange of data were assessed as follows: SAP, Business Objects, Oracle Data Integrator, Apatar, and for migration - SAXES.

All solutions showed strong positive features to their adhesion. But, for the OpenICGFw there arises the need for as yet inexistent resources not yet designed for the parameterization of these processes.

4.2 Notation adopted for Data Transformation in standard XML / GML

To convert the data, one solution explored refers to the copy of the original data for transformation generating a new set of data to enable the running of complex computational processes, presenting a specific notation in accordance with OGC [11] and W3C [44], achieving its implementation due to the need for automation of parallel processes, specifically for different purposes. Having seen many changes and adjustments of the same data set, to generate new datasets, then submitted to different data mining algorithms by means of the framework proposed.

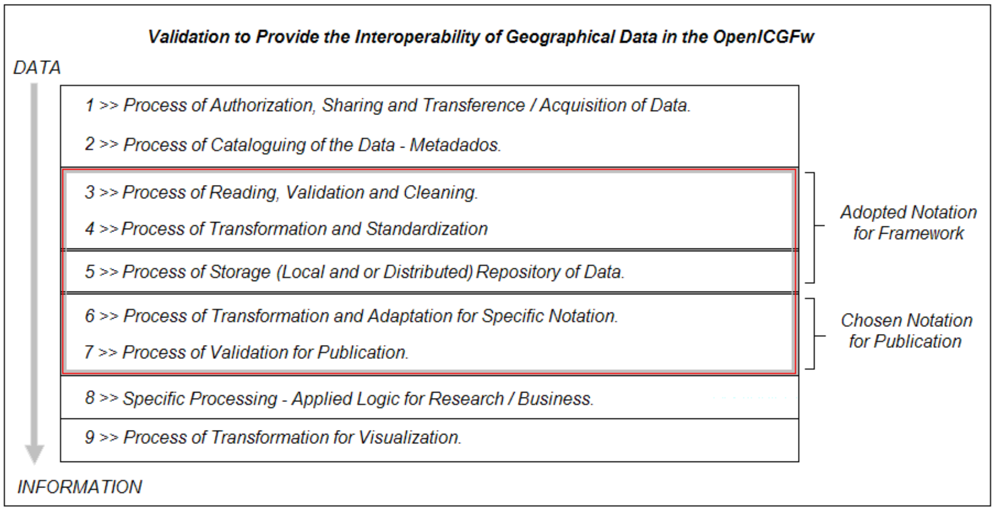

Validation to Provide the Interoperability of Geographical Data in the OpenICGFw
DATA
1 >> Process of Authorization, Sharing and Transference / Acquisition of Data.
2 >> Process of Cataloguing of the Data - Metadados.
3 >> Process of Reading, Validation and Cleaning.
4 >> Process of Transformation and Standardization
5 >> Process of Storage (Local and or Distributed) Repository of Data.
6 >> Process of Transformation and Adaptation for Specific Notation.
7 >> Process of Validation for Publication.
8 >> Specific Processing - Applied Logic for Research / Business.
9 >> Process of Transformation for Visualization.
Adopted Notation for Framework
Chosen Notation for Publication
INFORMATION

Figure 5: Architecture adopted to promote the interoperability of geographic data in the framework. Adapted from [3] Macário et. al. (2009) and [4] Pastorello et. al. (2009).

Regarding the storage of all data transmitted to a central archive, originally conceived by a freeNAS server (Network-Attached Storage) [45], as the best alternative because it is free software and especially for its connectivity through various communication protocols such as Samba, FTP, NFS and others.

Nevertheless, throughout this work and while running tests a surprising growth in processes concerning the handling of large volumes of data was detected, leading quickly to a limited capacity of the physical and logical structure. Therefore, consumption and handling of resources was restricted during the case study. This fact has triggered new thinking about other alternatives for working with distributed and also centralized data, leaving these assessments for future discussions on a better implementation and operation.

With the completion of the third, fourth and fifth process, illustrated in Figure 6, the result of the notation adopted for conversion of a project, identifying this notation via its ID, where the necessary metadata for cataloging in the base data framework is also collected. Among the main attributes we can refer to ID_Project, Project_Name, Format_Origin, Elaboration; these attributes are fundamental for the entire data conversion to be done well and properly screened later and vice versa. To accomplish the implementation of this notation in the framework, we designed a program in C++, containing specific features of a syntactic analyzer called Parser, responsible for reading and writing files with the XML/GML coding. This enabled it (Parser) to transform the data ready for handling in the following processes.

4.3 Exchange of Spatial Data

To promote the exchange of data, libraries available from the TerraLib project [39] were incorporated into the framework enabling the integration of specific resources for interoperability with products offered by the Brazilian National Institute for Space Research - INPE.

Already established free software was also incorporated, among them the GDAL (Geospatial Data Abstraction Library) project [46] due to various features being compatible with the framework. Through the GDAL/OGR (Simple Feature Library) project the exchange of raster and vector data was possible through the processes established to run the framework. Another important feature for using the GDAL refers to the large number of combinations that can be performed in the process of exchanging data.

Especially for the matrix data it was possible to reach a larger number of situations in the process of importing and transforming to different formats, enabling the conversion between 50 different formats (50x50) giving rise to 2500 possible combinations, another characteristic refers to 38 possible conversions for the creation and geo-referencing of data (38x38) allowing 1.444 different combinations. Originally both libraries require handling using command lines, although the sequence of these commands presents a specific syntax, a graphic interface has been made available for the data exchange module in the framework, adopting the MVC (Model View Controller) standard for the web.

This facilitates the running of processes for the conversion of the data with a friendly personalized interface for user navigation and especially to assist in the passage of parameters through the proposed framework. Specifically for the case study, the acquisition of the data submitted through the framework, it was possible to individually monitor each set. IBGE providing a volume of 551 megabytes (Mb) of data, 73 Mb from Epagri, 700 Mb from INPE, 60 Mb from the Local Government and 1.2 Gigabytes of data from the Ministry for the Environment. Featuring a total of 1.3 Gigabytes of data for the experiment distributed among the four sources selected for the experiment.

4.4 Promoting Collaboration in the Search for Information

For the flow of information, collaboration is detected as one of the important steps to minimize the lack of communication. But it is also important to note that cooperation takes place in different situations, only when it is possible to get all of the steps done, regardless of which information system is used.

For this, it is not enough just to be informed of the existence of certain (given) content, if it is not coordinated with its distribution. Only after the distribution and free access to (data) content, is collaboration made possible and consequently dynamic cooperation actually take place between the users as shown in Figure 6. The information flow for the proposed framework and the collaboration itself occurs with the participation of different users through a unique environment.

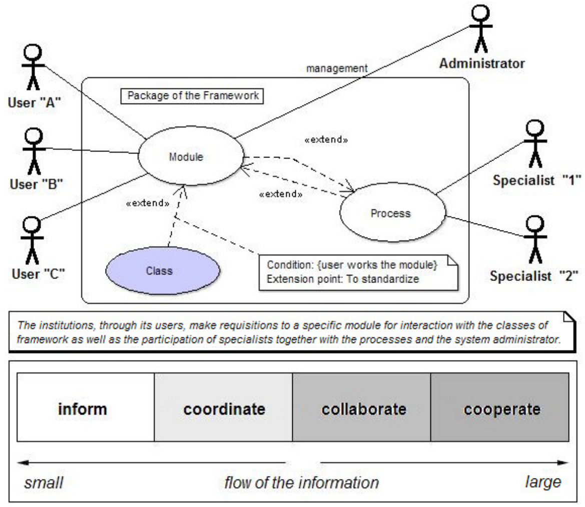

User "A"
User "B"
User "C"
management
Administrator
Package of the Framework
Module
«extend»
«extend»
«extend»
Process
Class
Condition: {user works the module}
Extension point: To standardize
Specialist "1"
Specialist "2"
The institutions, through its users, make requisitions to a specific module for interaction with the classes of framework as well as the participation of specialists together with the processes and the system administrator.
inform
coordinate
collaborate
cooperate
small
flow of the information
large

Figure 6: Intensity of information flow among users of the framework.

Therefore, to strengthen collaboration between different users and especially among public institutions, this initiative becomes an important means of communication and learning, and facilitates the integration [47] and cataloging of geographic data between different computer applications. Experiments were performed using the techniques of Association, Classification and Grouping [31].

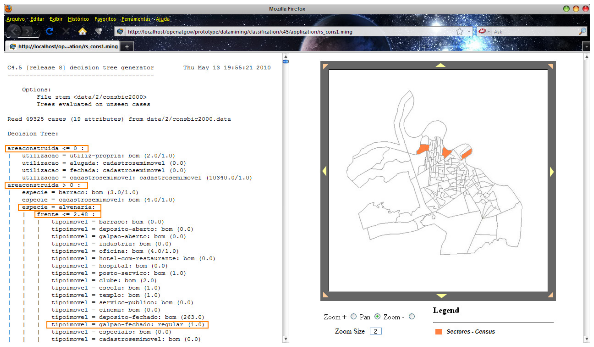
Also with the implementation of data mining, this framework seeks to expand the resources and assist in more complex analytical formulation of higher complexity in large spatial databases, as in the research in [48] and [49].

4.5 Implementing Data Mining

Using these techniques it was possible to generate rules and compose cartograms through automatic and semiautomatic computational methods. Initially, after the conversion of data as illustrated in (Section 4.2), for each mining technique the data set to be processed was standardized, this being necessary to go through the stages of selection, cleaning and making data appropriate for the classification process and the setting up of the decision tree.

Although for the technique of classification changes were made to the source code of the algorithm (C 4.5) and proposed in [50], adapting its implementation to run directly on the web browser to individually assess each rule found, significantly increasing the number of queries to understand situations that have previously gone undetected by other methods of data analysis. This has made it possible to assess and perfect the implementation of other algorithms, among them are those presented in [32] Apriori for association and k-means for clusters, as well as techniques for the investigation of spatial data for the web.

This experiment enabled the reduction of the existing complexity in the implantation of these processes of analysis. Figure 7 presents the scheme developed for the experiment with the C 4.5 algorithm. In this research adaptations were made to the original code to run on the Apache web server platform - (http://www.apache.org). After the assessment of the data obtained from the application of the C 4.5 algorithm it was possible to get satisfactory results which were also very significant for new experiments, especially aimed at large volumes of data.


Mozilla Firefox
Arquivo Editar Exibir Histórico Favoritos Ferramentas Ajuda
http://localhost/openatgcw/prototype/datamining/classification/c45/application/rs_cons1.ming
http://localhost/op...ation/rs_cons1.ming
Legend
Zoom + ○ Pan ◉ Zoom - ○
Zoom Size 2
Sectores - Census

```
C4.5 [release 8] decision tree generator       Thu May 13 19:55:21 2010
----------------------------------------

    Options:
        File stem <data/2/consbic2000>
        Trees evaluated on unseen cases

Read 49325 cases (19 attributes) from data/2/consbic2000.data

Decision Tree:

areaconstruida <= 0 :
|   utilizacao = utiliz-propria: bom (2.0/1.0)
|   utilizacao = alugada: cadastrosemimovel (0.0)
|   utilizacao = fechada: cadastrosemimovel (0.0)
|   utilizacao = cadastrosemimovel: cadastrosemimovel (10340.0/1.0)
areaconstruida > 0 :
|   especie = barraco: bom (3.0/1.0)
|   especie = cadastrosemimovel: bom (4.0/1.0)
|   especie = alvenaria:
|   |   frente <= 2.48 :
|   |   |   tipoimovel = barraco: bom (0.0)
|   |   |   tipoimovel = deposito-aberto: bom (0.0)
|   |   |   tipoimovel = galpao-aberto: bom (0.0)
|   |   |   tipoimovel = industria: bom (0.0)
|   |   |   tipoimovel = oficina: bom (4.0/1.0)
|   |   |   tipoimovel = hotel-com-restaurante: bom (0.0)
|   |   |   tipoimovel = hospital: bom (0.0)
|   |   |   tipoimovel = posto-servico: bom (1.0)
|   |   |   tipoimovel = clube: bom (2.0)
|   |   |   tipoimovel = escola: bom (1.0)
|   |   |   tipoimovel = templo: bom (1.0)
|   |   |   tipoimovel = servico-publico: bom (0.0)
|   |   |   tipoimovel = cinema: bom (0.0)
|   |   |   tipoimovel = deposito-fechado: bom (263.0)
|   |   |   tipoimovel = galpao-fechado: regular (1.0)
|   |   |   tipoimovel = especiais: bom (0.0)
|   |   |   tipoimovel = cadastrosemimovel: bom (0.0)
```

Figure 7: Assessment of the rules individualized through maps (cartograms).

A database with 49,325 tuplas was used, and after processing the algorithm it was possible to generate the decision tree. Apart from this, the processing of the algorithm enabled numerous mathematical classifications which are very difficult to be done and visualized manually by users. This is an essential requirement for improving the performance and the existence of ever more robust hardware computational resources. Also the reading itself of the decision tree presents a difficult interpretation of information, because, to facilitate understanding of the OpenICGFw, the rules are transformed into cartograms by means of queries using visual resources e contributing in the implementation of a webmapping application [34].

4.6 Processing and Visualizing the Information

The advantage of using a mechanism to integrate different sources of data, without any doubt it significantly reduces time destined for its users for elaboration and execution of research in search of precise information. However, on the other hand this integration requires exertion and organization which facilitate the dissemination of large volumes of data taking into consideration good practice in its use.

For this, in the OpenICGFw resources have been developed that allow the passage of parameters for processing in different data sources, the input of key parameters being solicited, specifically with respect to the origin of the sources for the preparation of data to be submitted to the search, parameters are also solicited to identify which algorithm should be adopted for the task of data mining, as well as the parameters that allow the customizing of the interface, adjusting the output for viewing. Before viewing the results, several processes are triggered after the execution of the query from the data archive of the framework. The transformation of different data occurs in parallel as in stage 6 and 7 (Figure 5) presented earlier.

Illustrated in Figure 8 is the processing of a query using the OpenICGFw, extending the process to other sources and returning the existing projects that cover the spatial dimension solicited and with different data formats, with extensions such as: .shp, .pdf, .tiff and .dwg. To better represent the output of all processes it is presented on screen in spatial format, making the process of exchange transparent between different projects for the end user.

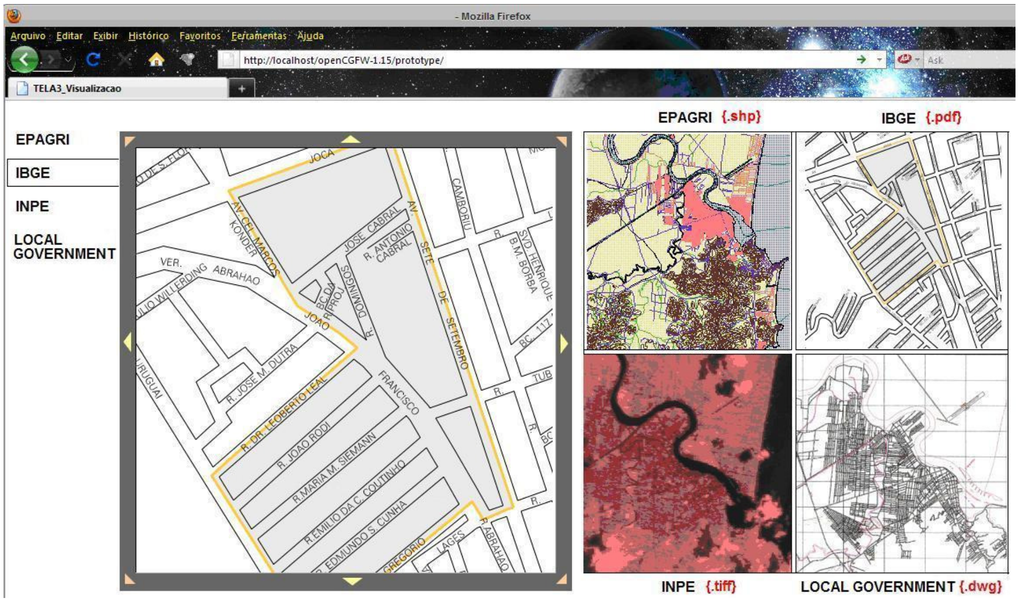


Figure 8: Screen of visualization after execution of the parameters.

Nevertheless, the possibility of identifying clusters that can contain and provide information about a region, mainly aims at assisting the reuse of existing projects, reducing costs related to data collection and development of new collaborative development.

5. CONCLUDING OBSERVATIONS

Given the structure of the fundamental points of view as well as the elementary ones, a need has been detected for regular monitoring by assessment of the tree with the profile of the impact of actions during the development process and throughout the life cycle of the software.

Regarding the interoperability of geographic data, some research presents conceptual schemes and standardization for the exchange. This leaves the development cycle out of the discussions. With the development of the geospatial strategic model of collaboration, the framework seeks to understand the different activities related to developing Webmapping (open source) solutions as well as understand the resources available in this software, enabling the reuse of existing functionalities. In addition, this paper mainly presents the initial stages of data collecting and planning of essential requirements for the construction of the development model of the framework proposed. Also with the study of assessment for the construction of the framework, it was possible to apply the Multicriteria Aid Decision Constructivist to identify the key elements that assist in the construction of the prototype.

This work also aims at detecting the evolutionary technological perspectives with the modernization and advancement and resources commonly used by public institutions in the treatment of geographic data, providing new expectations and job opportunities. Thus, it seeks to contribute to the technological advance to intensify the use of free standards and technologies.

With the development of a framework for free software, specific to the collaboration of spatial data it is intended to increase the involvement of a significant number of professionals in the adhesion of the free geo-technologies and consequently in the development of new functionalities for the framework. In addition to enabling new users to truly engage with the needs of institutions that devote a good part of their time in developing existing mechanisms or try to understand the code written by other developers.

By enabling the development of new mechanisms for collaboration and analysis of geographic data through the OpenICGFw framework it seeks to develop new research considering the spatial data mining allied the different variables involved in the most diverse segments of society, enabling, through this work, the origin of new collaborative projects that assist learning and knowledge application in managing spatial data for a group of actions.

ACKNOWLEDGEMENTS

Thanks go to the Graduate Program in Civil Engineering at University Federal of Santa Catarina, for the structure of the geo-engineering labs.

REFERENCES

[1] Steudler, Daniel; Törhönen, Mika-Petteri. 2010. Floss in Cadastre and Land Registration. Edition by Food and Agriculture Organization of the United Nations - FAO. Roma, Italy. (4-49). Available in: http://www.fig.net/pub/fao/floss_cadastre.pdf Accessed in: Jul. 2010.

[2] Câmara, Gilberto; Monteiro, M. Antonio; Paiva, A. João; Souza, C. Ricardo; Lopes, C. Júlio; Barbosa C. Cláudio; Freitas, M. Ubirajara. 2000. (in Portuguese) Intercâmbio de Dados Geográficos no Brasil: um formato aberto. INPE. São José dos Campos-SP, Brazil, 2000. Available in: http://www.dpi.inpe.br/geobr/ Accessed in: Jun. 2009.

[3] Macário, Geovana N. Carla; Medeiros, Bauzer, Claudia. 2009. Specification of a Framework for Semantic Annotation of Geospatial Data on the Web. SIGSPATIAL Special Volume 1(1): (Mar, 2009) 27-32. DOI = http://doi.acm.org/10.1145/1517463.1517466

[4] Pasttorello, Zonta Jr. Gilberto; Senra, Dias Arruda Rodrigo; Medeiros, Bauzer Claudia. 2009. A Standards-based framework to Foster geospatial data and process interoperability. Journal of the Brazilian Computer Society. Volume 15 (13-35).

[5] Ensslin, Leonardo; Montibeller, G. N.; Noronha, M. 2001. (in Portuguese). Apoio à decisão – Metodologia para Estruturação de Problemas e Avaliação Multicritério de Alternativas. Editora Insular. Florianópolis-SC, Brazil.

[6] Johnson, Bruce; Woolfolk, Walter; Miller, Robert; Johnson, Cindy. 2008. (in Portuguese). Projeto de software flexível: desenvolvimento de sistemas para requisitos variáveis. Editora LTC. Rio de Janeiro, Brazil.

[7] Santos, D Carlos. 2009. Open Source Software Projects Attractiveness, Activeness, and Efficiency as a Path to Software Quality: An Empirical Evaluation of Their Relationships And Causes. Doctoral Thesis. Southern Illinois University Carbondale. Available in: http://opensiuc.lib.siu.edu/dissertations/2/ Accessed in: Apr. 2010.

[8] Bana E Costa, C.A. 1993. (in Portuguese). Processo de apoio à decisão: actores e acções; estruturação e avaliação. Publicação CESUR, Volume 618 (31).

[9] eGOV - Governo Eletrônico, 2010. Available in: http://www.governoeletronico.gov.br - Accessed in: May. 2010.

[10] INDE – Infraestrutura Nacional de Dados Espaciais. 2009. Perfil de Metadados Geoespaciais do Brasil em Conformidade com a Norma ISO19115:2003. Available in: http://www.concar.ibge.gov.br/arquivo/Perfil_MGB_Final_v1_homologado.pdf Accessed in: May. 2010.

[11] OGC - Open Geospatial Consortium, Inc. 2009. Available in: http://www.opengeospatial.org - Accessed in: Apr. 2009.

[12] Horstmann, Cay. 2007. (in Portuguese) Padrões e Projeto Orientados a Objetos. 2ª. Edição – Editora Bookman – Porto Alegre-RS, Brazil.

[13] Silva, P. Ricardo. 2007. (in Portuguese) UML: Modelagem Orientada a Objetos. Editora Visual Books, Florianópolis-SC, Brazil, 2007.

[14] Pressman, S. Roger. 2006. Web Engineering: An Agile Discipline. XXI Simpósio Brasileiro de Banco de Dados e XX Simpósio Brasileiro de Engenharia de Software. Florianópolis-SC Brazil, 2006. Available in: http://www.rspa.com/download/Brazil2006.ppt Accessed in: Jul. 2010.

[15] Allen, Rob; LO, Nick; Brown, Steven. 2009. ZEND Framework in Action. Published by Manning Publications Co. Greenwich, CT.

[16] Freeman, Eric; Freeman, Elisabeth. 2007. Head First Desing Patterns. (second edition) – Publisher O´Reilly Media. Sebastopol – CA.

[17] Lisboa, G.S. Flávio. 2008. (in Portuguese). Zend Framework: Desenvolvimento em PHP5 orientado a objetos com MVC. Editora Novatec. São Paulo-SP, Brazil.

[18] Melo, Alexandre Altair de. 2007. (in Portuguese). PHP Profissional: Aprenda a desenvolver sistemas profissionais orientados a objetos com padrões de projeto. Editora Novatec. São Paulo-SP, Brazil.

[19] Mineto, Elton Luis. 2007. (in Portuguese). Frameworks para desenvolvimento em PHP. Editora Novatec. São Paulo-SP, Brazil.

[20] Soares, Wallace. 2009. (in Portuguese). Crie um Framework para sistemas web com PHP5 e Ajax. Edtiora Érica. São Paulo-SP, Brazil.

[21] Caratti, L. Ricardo; Silva, M. Leonardo. 2009. (in Portuguese). Joomla! Avançado: Aprenda a desenvolver componentes, módulos, plug-ins e templates para Joomla! usando PHP. Editora Novatec. São Paulo-SP, Brazil.

[22] Deboni, José Eduardo. 2007. (in Portuguese) Interoperabilidade: uma combinação eficiente de arquitetura e padrões. Revista infoGeo. Curitiba-PR, Brasil, Volume 50-(9).

[23] Sauvé, Jacques Philippe. 2007. O que é um framework? Available in:

http://www.dsc.ufcg.edu.br/~jacques/cursos/map/html/frame/oque.htm Accessed in: Mar. 2009.

[24] Goodchild, M; Egenhofer, J. Max; Fegeas, R.; Kottman, C. 1999. Interoperating Geographic Information Systems. Academic Publishers Springer. Volume 495 (30-110).

[25] Sherman, E. Gary. 2008. Desktop GIS Mapping the Planet with Open Source Tools. Publisher Pragmatic Bookshelf – Shelving Programming.

[26] Tsou, Ming-Hsiang. 2001. A Dynamic Architecture for Distributing Geographic Information Services on the Internet. Doctoral Thesis. University of Colorado at Boulder. Ph.D Geography.

[27] Rainer Simon, Peter Fröhlich. 2007. A mobile application framework for the geospatial web. Proceedings of the 16th international conference on World Wide Web, Banff, Alberta, Canada, (381-390).

[28] Tsalgatidou A., Pilioura, T.; 2002. An Overview of Standards and Related Technology In Web Services. International Journal of Distributed and Parallel Databases. Special Issue on E-Services. Publisher Springer Netherlands, Volume 12 (135-162).

[29] Fonseca, T. Frederico; Egenhofer, J. Max. 1999. (in Portuguese) Sistemas de Informação Geográficos Baseados em Ontologias. Articles in Refereed Journals - Cognition and Computation, Volume 1(155-180).

[30] Zarine Kemp, Lei Tan, Jacqueline Whalley. 2007. Interoperability for geospatial analysis: a semantics and ontology-based approach. ADC '07: Proceedings of the eighteenth conference on Australasian database – Volume 63 (83-92).

[31] Witten, Ianh; Frank, Eibe. 2005. Data mining: pratical machine learning tools and techniques (Second edition). Morgan Kaufmann Publishers. San Francisco-CA, USA. (37-76).

[32] Tan, Pang-Ning; Steinbach, Michael; Kumar, Vipin. 2009. Introduction to DATAMINING. Edition by Person Education, Inc.

[33] Casanova, Marco; Câmara, Gilberto; Davis, Clodoveu; Vinhas, Lúbia; Queiroz, Ribeiro de, Gilberto. 2005. (in Portuguese). Banco de Dados Geográficos. Editora MundoGEO. Curitiba-PR, Brasil. (10-85).

[34] Mitchell, Tyler; 2005. Web Mapping Illustrated. Using Open Source GIS Toolkits. Publisher O´Reilly Media Inc.(12-35: 59-209).

[35] Silva, P. Ricardo. 2007. (in Portuguese) UML: Modelagem Orientada a Objetos. Editora Visual Books, Florianópolis-SC Brazil.

[36] IBGE. Instituto Brasileiro de Geografia e Estatística. Geociências, 2009. Available in: http://www.ibge.gov.br/servidor_arquivos_geo/ Accessed in: Nov. 2009.

[37] Prefeitura Municipal de Itajaí – Mapas, 2010 Available in: http://www.itajai.sc.gov.br/mapas.php Accessed in: Feb. 2010.

[38] EPAGRI Empresa de Pesquisa Agropecuária e Extensão Rural de Santa Catarina – Mapas Digitais, 2010. Available in: http://ciram.epagri.sc.gov.br/mapoteca/ - Accessed in: Apr. 2010.

[39] INPE – Instituto Nacional de Pesquisas Espaciais. 2009. Softwares Livres - Tutorial de Programação TerraLib. Available in: http://www.terralib.org/docs/v313/ Accessed in: Feb. 2009.

[40] MMA - Ministério do Meio Ambiente do Brasil. Mapas Interativos – Geoprocessamento, 2009. Available in: http://www.mma.gov.br - Accessed in: Nov. 2009.

[41] Egenhofer M. J., Fegeas R., Goodchild M.F. 1997. Interoperating GISs Report of a Specialist Meeting Held under the Auspices of the Zarenius Project, Panel on Computational Implementations of Geographic Concepts. Santa Barbara, California. Available in: http://www.ncgia.ucsb.edu/conf/interop97/report.html Accessed in: Jun. 2010.

[42] Thakkar, Snehal; Knoblock, A. Craig; Ambite, L. Jose. 2007. Quality-Driven Geospatial Data Integration. Proceedings of the 15th International Symposium on Advances in Geographic Information Systems ACM GIS. New York, USA. DOI = http://doi.acm.org/10.1145/1341012.1341034

[43] GeoNetwork, 2010. Available in: http://geonetwork-openource.org/ Accessed in: Feb. 2010.

[44] W3C – World Wide Web Consortium, 2010. Available in: http://www.w3.org/ - Accessed in: May. 2010.

[45] FreeNAS - Project Open Source (Network-Attached Storage), 2010. Available in: http://freenas.org/doku.php/ - Accessed in: Feb. 2010.

[46] GDAL - Geospatial Data Abstraction Library – 2009. Available in: http://www.gdal.org/ - Accessed in: Feb. 2009.

[47] Tu, S.; Xu, L. ; Abdelguerfi M.; Ratcliff, J. 2002. Achieving interoperability for integration of heterogeneous COTS geographic information systems. Proceedings of the 10th ACM international symposium on Advances in geographic information systems ACM Press New McLean, Virginia, USA(162-67) DOI= http://doi.acm.org/10.1145/585147.585182

[48] Borgony, Vânia; Nievinski, Felipe; Bigolin, Nara. 2003. (in Portuguese) A Spatial Data Model for the Integration of Public Health Data. International SBC Workshop on Free Sofware. Porto Alegre-RS, Brazil, Volume 4 (119-122). Available in: http://www.inf.ufsc.br/~vania/WSL2003.pdf Accessed in: Jul. 2010.

[49] Bogorny, Vânia. 2006. Enhancing Spatial Association Rule Mining in Geographic Databases. Doctoral Thesis. Universidade Federal do Rio Grande do Sul. Programa de Pós-Graduação em Computação. Porto Alegre-RS, Brazil, 2006. Available in: http://www.inf.ufsc.br/~vania/publications.html Accessed in: Jul. 2010.

[50] Quinlan, J. Ross: 1993. C4.5 Programs For Machine Learning. Morgan Kaufmann Publishers, California. (17-25; 45-55).

[51] Fieldingo, T. Roy. 2000. Architectural Styles and the Design of Network-based Software Architectures. Doctoral Thesis. University of California, Irvine. (24-37).

[52] Macário, Geovana N. Carla; Santos, A. Jefersson; Medeiros, Bauzer, Claudia; Torres, S. Ricardo. 2010. Annotating data to support decision-making: a case study. Proceedings of the 6th Workshop on Geographic Information Retrieval. Zurich, Switzerland. DOI= http://doi.acm.org/10.1145/1722080.1722106

Geospatial Route Extraction from Texts

Euthymios Drymonas
Institute for the Management of
Information Systems/RC ATHENA
G.Bakou 17, 11524, Athens, Greece
+30 2106990522
edrimon@imis.athena-innovation.gr

Dieter Pfoser
Institute for the Management of
Information Systems/RC ATHENA
G.Bakou 17, 11524, Athens, Greece
+30 2106990522
pfoser@imis.athena-innovation.gr

ABSTRACT

The need to collect vast amounts of geospatial data is driven by the emergence of geo-enabled Web applications and the suitability of geospatial data in general to organize information. Given that geospatial data collection and aggregation is a resource intensive task typically left to professionals, we, in this work, advocate the use of information extraction (IE) techniques to derive meaningful geospatial data from plain texts. Initially focusing on travel information, the extracted data can be visualized as routes derived from narratives. As a side effect, the processed text is annotated by this route, which can be seen as an improved geocoding effort. Experimentation shows the adequacy and accuracy of the proposed approach by comparing extracted routes to respective map data.

Categories and Subject Descriptors

H.3.1 [INFORMATION STORAGE AND RETRIEVAL]: Content Analysis and Indexing - Indexing methods, Linguistic processing

General Terms

Algorithms

Keywords

user contributed geospatial data, data extraction, information extraction, geospatial data, natural language processing

1. INTRODUCTION

Geographic information, be it maps or 3D virtual worlds, are believed to be the future way for people to socialize, shop, and share information. In the foreseeable future, the map will become the interface of choice for the Internet [23]. In an increasing number of (Web) applications space is however not only used as metadata to structure and access information, but also as the actual content resource. Overall, the most significant advantages of geospatial data are its (i) unambiguous nature, i.e., categories and keywords are up for interpretation, a geographic coordinate is not, and (ii) the simplicity of the matching interface, i.e., maps.

To “geo enable” the Web, two major issues need to be addressed. One, content needs to be related to geographic co-ordinates, i.e., geocoded. Second, sufficient amounts of geospatial data need to be available, e.g., (road) networks, address information, POIs, routes, etc. *Geocoding* has been exhaustively addressed not only in literature but also by a series of products (cf. Google Maps API [11]), all sharing the basic approach of comparing text strings to gazetteer entries that are linked to coordinates. A different issue is the availability of *sufficient geospatial data sets* for any types of application. Here, with the proliferation of the Internet as the primary medium for data publishing and information exchange, we have seen an explosion in the amount of online content available on the Web. Thus, in addition to professionally-produced material being offered free on the Internet, the public has also been allowed, indeed encouraged, making its content available online to everyone by means of *user-contributed content*. The aim is to harness the ability humans have to massively collect and share knowledge with the ultimate goal of digitizing the world (from a geospatial point of view). As early maps were traces of people’s movements in the world, i.e., view representations of people’s experiences, digitizing the world in this context relates to collecting pieces of knowledge gained by a human individual tied not only to space and time, but also to her context, personal cognition, and experience. Through *intentional* (e.g., narratives, geo-wikis, geocoding photos) or *unintentional effort* (e.g., routes from their daily commutes), simple users create vast amounts of data concerning the real world that contain significant amounts of information. The ambitious aim in such a crowdsourcing effort will be however to go beyond purposefully contributed data and to *include any type of available content such as existing Web pages in the data collection effort*. This potentially vast amount of data will lead to a digitized world beyond mere collections of co-ordinates and maps.

Of importance to both, the geocoding and the crowdsourcing approach, is an *understanding of textual content with respect to the geospatial data that it contains*. To this respect, this work proposes an Information Extraction (IE) approach based natural

ACM SIGSPATIAL DMG Workshop, Nov 2, San Jose, CA, U.S.A.
Copyright 2010 ACM ISBN 978-1-4503-0430-6/10/11...$10.00."

language processing (NLP) techniques towards the understanding, detection and extraction of geospatial data nuggets from texts. Using text engineering methods, we propose for the context of travel information a (extendable) set of rules that allows us to detect travel information in written texts. This rule base can be extended; however, our experimentation showed that after considering a certain number of documents an accurate detection of route information in newly added texts can be achieved. We combined this approach with geocoding and routing functionality to derive actual route information. As such the proposed approach is a hybrid system incorporating, both, aspects of geocoding, i.e., texts are not only related to location information but to actual routes, and geospatial data extraction, i.e., actual route information is extracted from texts without having any prior knowledge. An empirical evaluation using actual texts from travel guides and travel diaries shows the usefulness and accuracy of the proposed approach.

Related work exists towards two general directions, (i) geocoding and (ii) extraction of routes from texts. With respect to geocoding, we can exemplary cite [16], one of the first works on geocoding and describing a navigational tool for browsing web resources by geographic proximity as an alternative means for Web navigation. Web-a-Where [1] is another system for geocoding Web pages. It assigns to each page a geographic *focus* — a locality that the page discusses as a whole. The tagging process targets large collections of Web pages to facilitate a variety of location-based applications and data analyses. The approach presented in [3] proposes the use of the Web's geographic information to populate address databases, i.e., parse Web pages for useful address information and populate an address database with the available information. The work presented in [13] is identifying and disambiguating references to geographic locations. A method for calculating the geographic breadth of a Web page is given in [9]. Another method that uses information extraction techniques to geocode news is described in [22]. In the realm of geocoding, a range of related *commercial products* exist. Google Maps API provides geocoding services [11]. A similar service is Yahoo Yellow Pages [17]. What is common to those services is that they simple try to geocode a given input string. MetaCarta on the other hand [15] provides tools and services that also geoparse and then geocode text content using natural language processes and highly refined geodata. The approach in [12] uses state-of-the-art tools in their work for extracting geographical information from data. The results can be used for geographic search on the Web, in GIS applications, for categorizing documents, etc. However, no evidence is given to the existence of a route extraction mechanism. In this context, we can cite [18], which aims at mapping natural language descriptions to a custom-created sidewalk database, i.e., this approach is not generally applicable to arbitrary routes since developed in a controlled environment and limited vocabulary. Work towards the classification of route-relevant expressions is presented in [26]. However, no actual routes are produced. [7] aims at extracting a transportation network graphs from Web documents. Using a given set of seed locations, Web documents are retrieved to identify candidate transportation nodes between the locations.

The outline of the remainder of this work is as follows. The contribution, namely geospatial data extraction from texts using information extraction methods is detailed in Section 2. Based on this approach, Section 3 presents an experimental evaluation that focuses on route extraction from texts. Finally, Section 4 gives conclusions and directions for future research.

2. GEOSPATIAL DATA EXTRACTION FROM TEXTS

In our approach, we apply Information Extraction (IE) methods for deriving geospatial content from narratives. IE is generally defined as the process of locating user-specific information in electronic documents, "*the name given to any process which selectively structures and combines data which is found, explicitly stated or implied, in one or more texts"* [5]. In the present effort our focus will be on content that contains rich geospatial data such as travel literature.

Given travel guides and travel diaries, our objective is to correctly recognize location and direction information so as to construct actual route datasets that can be visualized on a map.

2.1 Overview

A precursor to extracting route information from texts and to actually construct a map, is to extract a meaningful and coherent series of points that describe the narrated route.

There is a huge amount in the WWW of texts regarding spatial content, like travel blogs or travel diaries and guides that contain rich information that could be exploited and organized in automatic ways. Instead, apart from the fact that users contribute their own narratives every day, these documents are not analyzed by computers in order to exploit semantic information and are treated as bags of words. Our method makes use of state-of-the-art Information Extraction methods to derive meaningful information by analyzing such free text narratives, extracting names of places as well as relative information between them. In this work we extract information like "head north for 20 meters and meet Key bar". We extract relative and absolute information regarding a place. This information would reveal places that cannot be geocoded (for example "Key bar" that a geocoder can't recognize), but mentioned explicitly in a text narrative. Thus a main advantage of our method is that we use only linguistic, semantic and contextual information contained in free text narratives, without making use of supervised methods (e.g., gazetteers, lists) in order to extract meaningful named entities (i.e. places) and relations between them.

By using IE techniques, we also try to bypass the important problem of *ambiguity*, i.e., not falsely linking identifiers to coordinates such as when the name of a geographic location shares a non-geographic meaning as well (George Washington vs. Washington DC) or distinct geographic locations share the same name (London, England vs. London, Ontario). Disambiguation of geographic entities is achieved by properly identifying the context of the identifier in a sentence. In addition, IE techniques help in addressing the problem of incomplete gazetteers and place name variations and abbreviations.

The IE system used in this work consists of three principal parts, (i) the linguistic pre-processing part, (ii) the document IE semantic analysis part (the core feature extraction process) and

(iii) the geocoding part. The various system components and relationships are shown in Fig. 1. The system has been implemented as a pipeline application of individual tools using the GATE - General Architecture for Text Engineering platform [6], a software framework for natural language processing and engineering. GATE allows for the embedding of different types of language resources (ontologies, lexicons, etc.) and modules that perform various types of processing in the form of plugins (CREOLE components). Each component has to be implemented as a Java Bean with a well defined input/output interface. Furthermore GATE provides a convenient graphical interface for developing and/or evaluating components for various natural processing tasks (cf. the use of this interface in visualizing annotations in Section 3). For a specific processing task, an arbitrary number of components may be used sequentially in what is termed a processing *pipeline*.

In what follows, we describe in necessary detail the processing pipeline, which overall uses a document in plain text as input and, as shown in Section 3, produces a map in the form of a KML file [19] that can be viewed by means of, e.g., Google Earth.

2.2 Linguistic pre-processing tools

Linguistic pre-processing tools analyze natural language documents in terms of words, sentences, part-of-speech and morphology. We selected the ANNIE tools, contained in the GATE release, to perform this initial part of analysis. To this task, our processing pipeline comprises of a set of four modules: (i) the ANNIE tokeniser, (i) the (ANNIE) Sentence Splitter, (iii) the ANNIE POS Tagger and (iv) the WordNet Lemmatiser.

The intermediate processing results are passed on to each subsequent analysis tool as GATE document annotation objects. The output of this analysis part is the analyzed document in CAS/XML format, an XML scheme called Common Annotation Scheme allowing for a wide range of annotations, structural, lexical, semantic and conceptual [21]. This document is temporarily stored in the system, so as to be accessed by the subsequent CAFETIERE semantic analysis component. CAFETIERE combines the linguistic information acquired by the pre-processing stage of analysis with knowledge resources information, namely the lookup ontology and the analysis rules to semantically analyse the documents and recognize spatial information.

The first step in the pipeline process is *tokenisation,* i.e., recognising in the input text basic text units (tokens), such as words and punctuation and *orthographic analysis*, i.e., the association of orthographic features, such as capitalisation, use of special characters and symbols, etc. to the recognised tokens [25]. The tools used are ANNIE Tokeniser and Orthographic Analyser. The ANNIE tokenizer distinguishes five types of tokens: *word, number, symbol, punctuation* and *space* tokens. The orthographic analysis process of the tool is paired with tokenisation analysis rule-based processing and distinguishes four orthographic categories for the respective token types: *upperInitial, allCaps, lowercase* and *mixedCaps* categories. These token types will be used in the rule-based CAFETIERE IE engine for recognizing placenames and spatial relations.

Sentence splitting, in our case the ANNIE sentence splitter aims at the identification of sentence boundaries in a text. Though a seemingly trivial task, sentence splitting can become quite complex due to the ambiguous or dual function of certain punctuation marks. A dot, for example, may indicate both an abbreviation and a sentence end and, among other uses, it can also be employed in acronyms and as indicator of decimal digits of a real number.

Part-of-speech (POS) tagging is the process of assigning a part-of-speech class, such as Noun, Verb etc. to each word in the input text. The ANNIE POS Tagger implementation is a variant of Brill Transformation-based learning tagger, which applies a combination of lexicon information and transformation rules for the correct POS classification.

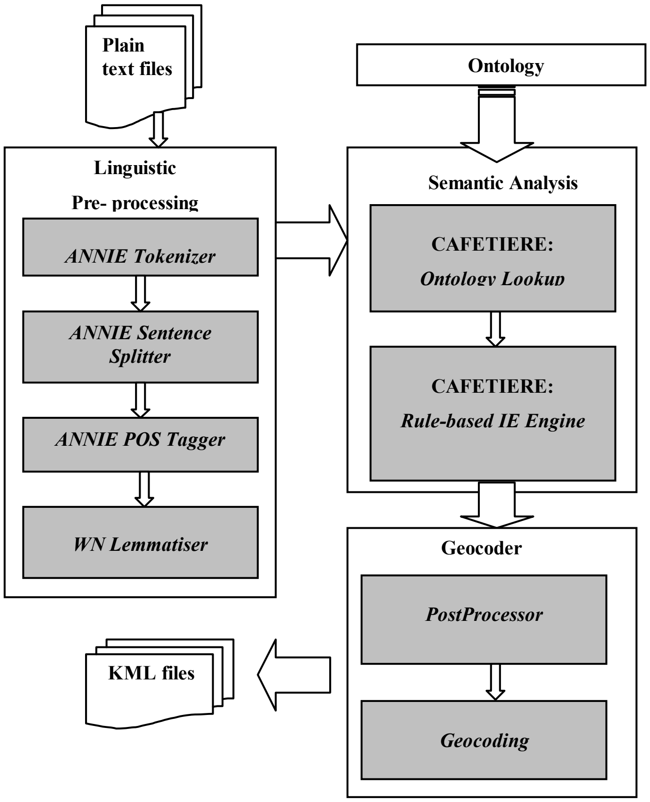


Fig. 1. GATE pipeline

Lemmatisation is used for text normalisation purposes. With this process we retrieve the tokens' base form e.g., for words: ["travelling", "traveler", "traveled"], "are", the corresponding lemmas are: "travel", "be". We exploit this information in the semantic rules section. For this purpose we implement the JWNL WordNet Java Library API [8] for accessing the WordNet relational dictionary. The output of this step is included it in GATE document annotation information.

2.3 Semantic Analysis Tools

Semantic analysis relates the linguistic processing results to ontology information and applies analysis rules, i.e., *documents* *are analyzed semantically to discover spatial concepts and relations*.

CAFETIERE [2] is a standalone system combining linguistic pre-processing and relevant for our work, semantic analysis. The CAFETIERE Information Extraction Engine module objective is to compile the set of the semantic analysis grammar rules in a cascade of finite state transducers so as to recognise in text the concepts of interest. For this purpose the CAFETIERE IE Engine combines all previously acquired linguistic and semantic (lookup) information with contextual information. We modified CAFETIERE to process documents in a GATE pipeline and perform only ontology lookup and rule-based semantic analysis. The input to this process are the GATE annotation objects resulting from the linguistic pre-processing stage stored in CAS/XML format for each individual document.

2.3.1 Cafetiere Ontology Lookup

The CAFETIERE Ontology lookup module accesses a previously built ontology to retrieve potential semantic class information for individual tokens or phrases. All types of conceptual information, related to domain specific entities, such as terms or words in general that denote spatial concepts or properties and relations of domain interest are pre-defined in this ontology, built by an expert. The ontology used in our experimentation was created by manually analyzing a large number of texts and iteratively refining the ontology so as to automatically produce results that are close to what an expert user would have manually created.

Consider the partial ontology shown in Fig. 2. Class “LOCVERB” stores verbs that when matched to a text phrase are likely to indicate a spatial relationship between the corresponding referenced concepts. For example, in the phrase “*cross over the bridge and head to Fifth Avenue*”, the existence of words contained in “LOCVERB” class denoting spatial information, like “*cross*” and “*head to*”, help us derive the desired information. Approximating the human notion when building phrases, we are extracting “Fifth Avenue” as a desired place name from this sample sentence.

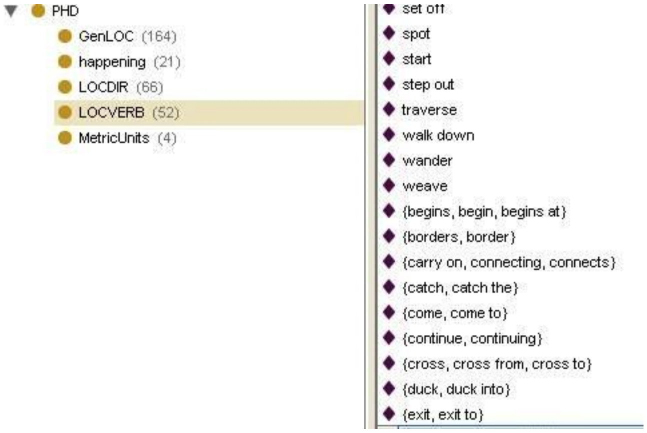


Fig. 2. Sample ontology contents (Protégé ontology editor)

With the application of semantic rules that make use more characteristics of the language, like part-of-speech or orthography (e.g., named entities are written uppercase), as we will see in Section 2.3.2, we are extracting the wanted features from text. Also, the results of this process do not include at this stage any information regarding place names extraction. The subsequent application of the semantic analysis rules undertakes the tasks of disambiguation and the extraction of spatial information. The lookup ontology consists of OWL statements. We can easily add or remove semantic classes or their respective instances by using an ontology editor such as Protégé [20] as shown and used in the example of Fig. 2.

2.3.2 Cafetiere Rule-Based IE engine

The semantic analysis rules, based on CAFETIERE specifications are developed as a single set of context-sensitive/context-free grammar (CSG/CFG) rules.

The CAFETIERE Information Extraction Engine module objective is to compile the set of the semantic analysis grammar rules in a cascade of finite state transducers so as to recognise the concepts of interest in plain texts. For this purpose the CAFETIERE IE Engine combines all previously acquired linguistic and semantic (lookup) information with contextual information found in the plain texts. The semantic analysis rules, are developed as a set of context-sensitive/context-free grammar (CSG/CFG) rules.

An example of a CAFETIERE rule formalism is as follows:

```
[s=__x, target=__trglabel,
rulid=relation8]=>
\
[lookup="LOCVERB", pos=VB, token=__x],
[lookup="LOCDIRECTION", token=__x],
[pos=IN, token=__x]?,
[pos=DT, token=__x]?,
[orth=uppercase, token=__trglabel,
token=__x]{1,4},
[lookup="GenLOC",
tokentoken=__trglabel, token=__x]?
/
```

In this rule formalism, the left part of the rule, before the arrow symbol (=>) is called left-part side of the rule (LHS), while the part appearing after the arrow symbol is called right-hand side of the rule (RHS). Each constituent of the RHS is in the form of single minimal textual units where words in the sentence are matched, while the LHS describes features where the final extracted text spans will be held. In our specific sample rule, LHS contains the rule’s id and two features, *s* and *target*, where we store the final information. For the above sample rule to be applied, the sentence snippet that should be matched should start with a verb matched in the lookup ontology as a verb denoting spatial information, the immediate next token should be a word showing directional information (ex. north, south), followed by a token with a part-of-speech tag of *IN* or *DT* (i.e. *preposition/subordinating conjunction* or *determiner*, as defined in [14][1]. The rule formalism provides both standard iteration (?,

[1] Example site with penn tagset: http://www.mozart-oz.org/mogul/doc/lager/brill-tagger/penn.html

+, *) and iteration range operators (e.x. in the above rule {1,5} means 1 to 5 times of consecutive uppercase tokens). The output placename will be written in the LHS entity reference feature named *target*.

As an example text snippet, let us consider the following example: "*From the tower, head east along the Amstel river to take in the* ...". The rule above specifies a pattern where firstly a token (i.e., "*head*"), matching the ontology class "LOCVERB" is extracted as an instance of the respective class, and denoting a verb that could be expressing spatial information. This token is recognized by the POS tagger as verb, so it also matches the required rule POS feature "VB". In the same way, the other tokens are recognized, with respect to their POS tag or their appearance in the lookup ontology. For example the POS tags *"in"* (preposition/subordinating conjunction) and *DT* (determiner) are matching the tokens "*along*" and "*the*", respectively. Finally, a token with an orthography typical for proper names (i.e., *uppercase*) is matched and since it co-occurs in a sentence with the other rule constituents, it is recognized as a spatial object.

In conclusion, the incremental variable *__trglabel,* attached in the *target* feature of the rule gets the value "*Amstel river*". By incremental variable we mean that the matched tokens after each one matched rule constituent are kept into this variable. Similarly, we capture in the *s* feature the contents of incremental variable *__x*, which is the phrase "*head east along Amstel river*". For more information about the CAFETIERE rule formalism, the reader is referred to [2]. The phrase "*Amstel river*" will be kept for geocoding by the Geocoding pipeline module, while the phrase "*head east along Amstel river*" kept in the *s* feature will be annotated visually in the GATE platform by the PostProcessor pipeline module (cf. Fig. 3).

The output of CAFETIERE is stored in a CAS/XML file, which for this example, is as follows:

```
<tok id="t211" pos="VB" lem="head"
   lookup="LOCVERB"
   orth="lowercase">head</tok>
<tok id="t212" pos="JJ" lem="east"
   lookup="LOCDIR" orth="lowercase"
   >east</tok>
<tok id="t213" pos="IN" lem="along"
   lookup="NIL"
   orth="lowercase">along</tok>
<tok id="t214" pos="DT" lem="the"
   lookup="NIL"
   orth="lowercase">the</tok>
<tok id="t215" pos="NNP" lem="amstel"
   lookup="NIL"
   orth="upperInitial">Amstel</tok>
<tok id="t216" pos="NN" lem="river"
   lookup="GenLOC"
   orth="lowercase">river</tok>
<tok id="t217" pos="TO" lem="to"
   lookup="NIL"
   orth="lowercase">to</tok>
<tok id="t218" pos="VB" lem="take"
   lookup="LOCVERB"
   orth="lowercase">take</tok>
<tok id="t219" pos="IN" lem="in"
   lookup="NIL"
   orth="lowercase">in</tok>
<Prelation id="pr2"
   label="head east along amstel
   river"
   source="" target="Amstel river"
   rulid="relation8" tokrefs="t211
   t212 t213 t215 t216" />
```

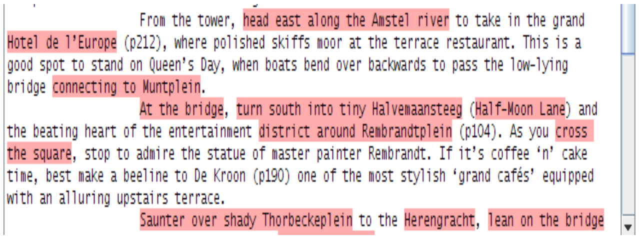
Note that ids like "t211" were assigned by GATE to each token in the previous pre-processing step and they are kept in feature "*tokrefs*".

The rules are stored in a plain text file, which is read during the initialization of the CAFETIERE module, thus allowing us to easily provide our system with new rules.

In the following sections, we describe the two modules that follow the semantic analysis process, namely the Postprocessor and the Geocoding module.

2.4 Postprocessor

The Postprocessor collects the output results of semantic analysis from the CAS/XML and relates it to the original text. Using the Castor tool [4], this module passes the token ids back to GATE to create annotation sets for the actual documents examined. Fig. 3 shows such a sample annotation for walk descriptions in a travel guide (cf. content used in experiments of Section 3). The Postprocessor module then passes the results (like "Amstel River" from the previous example) to the Geocoder.



From the tower, head east along the Amstel river to take in the grand Hotel de l'Europe (p212), where polished skiffs moor at the terrace restaurant. This is a good spot to stand on Queen's Day, when boats bend over backwards to pass the low-lying bridge connecting to Muntplein.

At the bridge, turn south into tiny Halvemaansteeg (Half-Moon Lane) and the beating heart of the entertainment district around Rembrandtplein (p104). As you cross the square, stop to admire the statue of master painter Rembrandt. If it's coffee 'n' cake time, best make a beeline to De Kroon (p190) one of the most stylish 'grand cafés' equipped with an alluring upstairs terrace.

Saunter over shady Thorbeckeplein to the Herengracht, lean on the bridge

Fig. 3. Original document annotated with extracted content

2.5 Geocoding and Routes

The semantic analysis provided us (among others) with *place name information*, i.e., the place name identifiers contained in the text, e.g., Amstel river, Muntplein, etc. To determine their actual location, these identifiers need to be geocoded. For this task we rely on the open-source module GeoGoogle [10], a Java API utilizing the Google Geocoder service, which is part of the Google Maps API [11].

The retrieved results are of varying accuracy. In the experiments of Section 3, only results of GeoAddressAccuracy >= 5 (cf. [11]) were used. This value corresponds to "street level accuracy", i.e., somewhere on a specific road. In addition, spatial outliers are detected by calculating the distance between sets of points, i.e., if a retrieved geocoding result would extend a path by more than *x* km it is omitted. For the case of the city guides used in Section 3, *x* was chosen to be 1km.

To then retrieve a *route*, the filtered geocoded place marks need to be connected so as to create a valid road path. In order to tackle this problem, we implemented a Java wrapper for directions feature of the Google Maps API (i.e., a wrapper similar to GeoGoogle for routing). This wrapper allows us to compute a shortest-path between place marks using the Google street network data. The result comprises a polyline for the route to follow in the road network. This is the final step of our pipeline implementation, with the geocoder module creating respective KML files of the respective routes. The following section showcases this approach and gives actual routes from example datasets.

2.6 Summary

Focussing on texts that contain route information, we use an information extraction approach that utilizes a location ontology to describe spatial relationships and properties in combination with a rule-based IE engine to extract place and, connecting them in sequence, route information. A main advantage of our system is that we *do not rely on exhaustive gazetteer lists*, but a *relatively small in size ontology to annotate texts and extract geospatial data*.

The following section gives some specific examples that show the applicability of the approach.

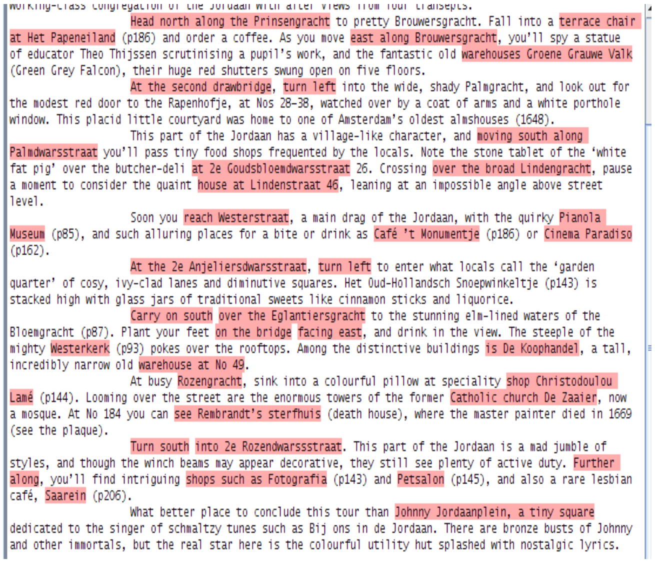
3. Experimental Evaluation

The following experimental evaluation tries to assess the quality of the proposed approach by comparing textual route descriptions with their actual map counterparts. For this purpose, we used content from actual Lonely Planet travel guides (Amsterdam, Budapest and Melbourne). In those guides walking tours are given by means of (i) a textual description and (ii) an accompanying map. Our objective was to recreate the map by processing the textual description with the approach advocated in Section 2, i.e., extract place information and geocoded as many as possible to actual show the created route on a map. The results are given in the following figures, which show a Google Earth visualization of the resulting KML next to the original travel guide map (Fig. 5).

3.1 Travel Guides and Routes

A complete route extraction example is shown in Fig. 4 (© Lonely Planet, Amsterdam City Guide – content used under "Fair Use" terms). Fig. 4(a) shows the annotated text of the guide after being processed by the IE system. Placemarks and movement information is highlighted. Fig. 4(b) visualizes the route extracted from the annotated text after being processed by the geocoder and routing engine. When compared to the original route that accompanied the text in Fig. 4(b), the two routes, although not an exact match are very similar.

Table 1 gives an overview of the actual text sizes and annotation results of the various city guides used in this experimentation and the respective processing results. For example, the text as shown partially in Fig. 4(a) comprises 520 words and 38 phrases were annotated, i.e., marked as containing place names or other relevant spatial information. Out of those annotations, 25 were actual place names and using GeoGoogle, we were able to geocode 10 entries. The resulting route is shown in Fig. 4(c). Respective numbers are given for the other three case studies. It is worth mentioning that *the quality of the resulting route highly depends on the geocoding tool as in the Amsterdam example, only 10 out of 25 recognized place names were geocoded.* Nevertheless, the produced result resembles the original route to a very large degree.


working-class congregation of the Jordaan with alter views from four transepts.

Head north along the Prinsengracht to pretty Brouwersgracht. Fall into a terrace chair at Het Papeneiland (p186) and order a coffee. As you move east along Brouwersgracht, you'll spy a statue of educator Theo Thijssen scrutinising a pupil's work, and the fantastic old warehouses Groene Grauwe Valk (Green Grey Falcon), their huge red shutters swung open on five floors.

At the second drawbridge, turn left into the wide, shady Palmgracht, and look out for the modest red door to the Rapenhofje, at Nos 28–38, watched over by a coat of arms and a white porthole window. This placid little courtyard was home to one of Amsterdam's oldest almshouses (1648).

This part of the Jordaan has a village-like character, and moving south along Palmdwarsstraat you'll pass tiny food shops frequented by the locals. Note the stone tablet of the 'white fat pig' over the butcher-deli at 2e Goudsbloemdwarsstraat 26. Crossing over the broad Lindengracht, pause a moment to consider the quaint house at Lindenstraat 46, leaning at an impossible angle above street level.

Soon you reach Westerstraat, a main drag of the Jordaan, with the quirky Pianola Museum (p85), and such alluring places for a bite or drink as Café 't Monumentje (p186) or Cinema Paradiso (p162).

At the 2e Anjeliersdwarsstraat, turn left to enter what locals call the 'garden quarter' of cosy, ivy-clad lanes and diminutive squares. Het Oud-Hollandsch Snoepwinkeltje (p143) is stacked high with glass jars of traditional sweets like cinnamon sticks and liquorice.

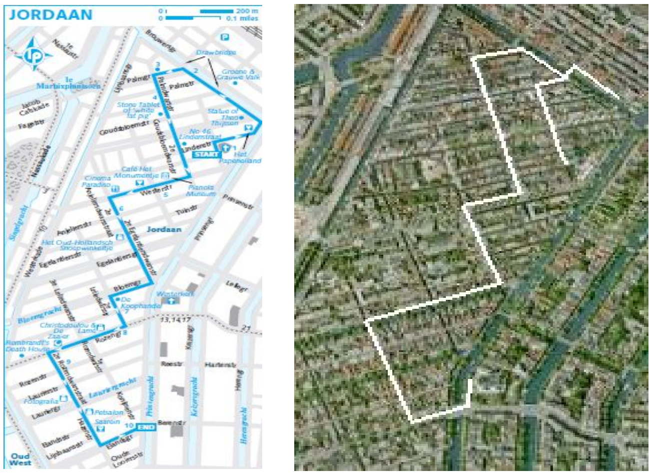
Carry on south over the Eglantiersgracht to the stunning elm-lined waters of the Bloemgracht (p87). Plant your feet on the bridge facing east, and drink in the view. The steeple of the mighty Westerkerk (p93) pokes over the rooftops. Among the distinctive buildings is De Koophandel, a tall, incredibly narrow old warehouse at No 49.

At busy Rozengracht, sink into a colourful pillow at speciality shop Christodoulou Lamé (p144). Looming over the street are the enormous towers of the former Catholic church De Zaaier, now a mosque. At No 184 you can see Rembrandt's sterfhuis (death house), where the master painter died in 1669 (see the plaque).

Turn south into 2e Rozendwarssstraat. This part of the Jordaan is a mad jumble of styles, and though the winch beams may appear decorative, they still see plenty of active duty. Further along, you'll find intriguing shops such as Fotografia (p143) and Petsalon (p145), and also a rare lesbian café, Saarein (p206).

What better place to conclude this tour than Johnny Jordaanplein, a tiny square dedicated to the singer of schmaltzy tunes such as Bij ons in de Jordaan. There are bronze busts of Johnny and other immortals, but the real star here is the colourful utility hut splashed with nostalgic lyrics.

(a)



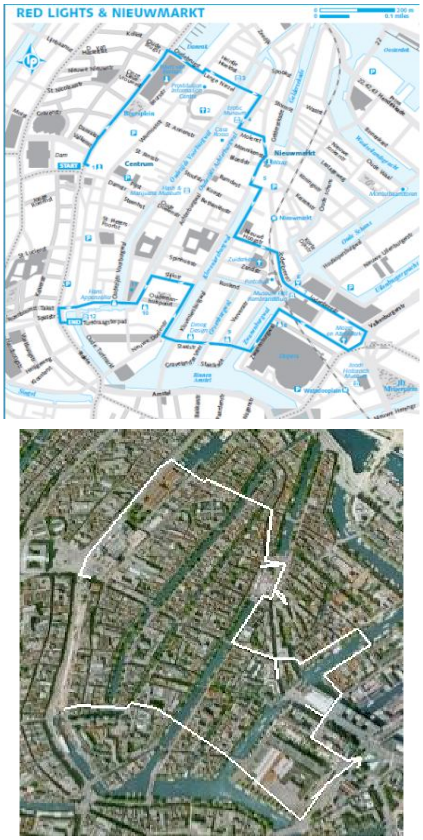
(b) (c)

Fig. 4. Amsterdam – route extraction example

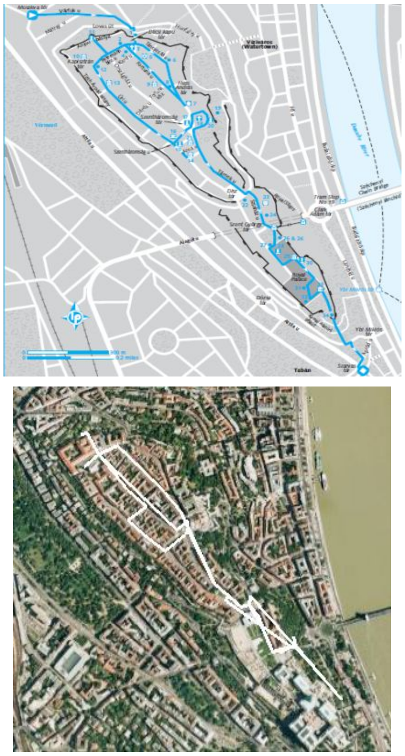
Further route extraction examples are shown in Fig. 5, including another route for Amsterdam, plus routes for Budapest, Hungary and Melbourne, Australia. Although in each case not all place names identified in the text were geocoded, each route clearly resembles the original one shown by means of a respective map. With better geocoding algorithms, which are beyond the scope of this work, the obtained route results could be considerably improved.

Table 1. Texts and processing results

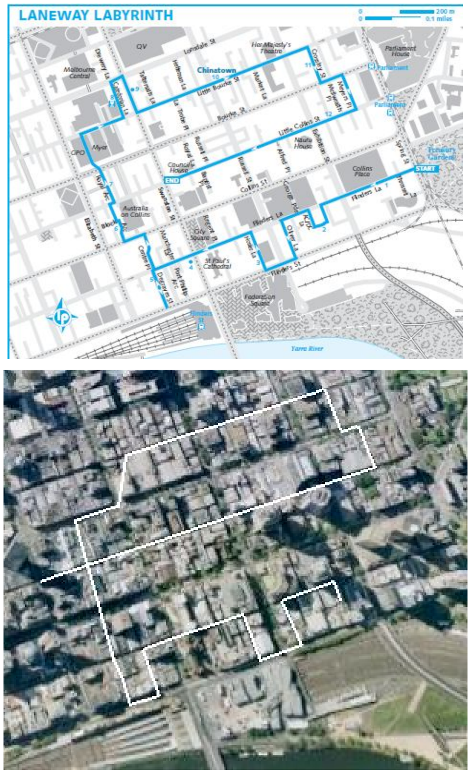
	Amsterd am/ Jordaan	Amster-dam – Nieuw-markt	Buda-pest	Mel-bourne
Nof. words total	520	566	1625	647
Annotations found	38	42	101	47
Place names found	25	32	58	35
Geocoded Place names	10	16	21	18
Geocoding percentage	*40%*	*50%*	*36%*	*51%*



(a) Amsterdam



(b) Budapest

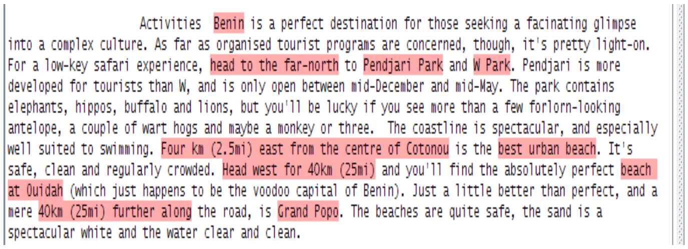


(c) Melbourne

Fig. 5. Route extraction examples (© Lonely Planet, City Guides)

3.2 Narratives and Routes

Fig. 6 shows a somewhat different example in the form of a travel diary containing narrative about a trip to Benin, Africa. The extracted route information and KML visualization are shown respectively. Please note that while more text portions have been identified, geocoding failed due to a lack of gazetteer data. This example should illustrate that our proposed approach is universally applicable and produces results for various types of content.



Activities Benin is a perfect destination for those seeking a facinating glimpse into a complex culture. As far as organised tourist programs are concerned, though, it's pretty light-on. For a low-key safari experience, head to the far-north to Pendjari Park and W Park. Pendjari is more developed for tourists than W, and is only open between mid-December and mid-May. The park contains elephants, hippos, buffalo and lions, but you'll be lucky if you see more than a few forlorn-looking antelope, a couple of wart hogs and maybe a monkey or three. The coastline is spectacular, and especially well suited to swimming. Four km (2.5mi) east from the centre of Cotonou is the best urban beach. It's safe, clean and regularly crowded. Head west for 40km (25mi) and you'll find the absolutely perfect beach at Ouidah (which just happens to be the voodoo capital of Benin). Just a little better than perfect, and a mere 40km (25mi) further along the road, is Grand Popo. The beaches are quite safe, the sand is a spectacular white and the water clear and clean.

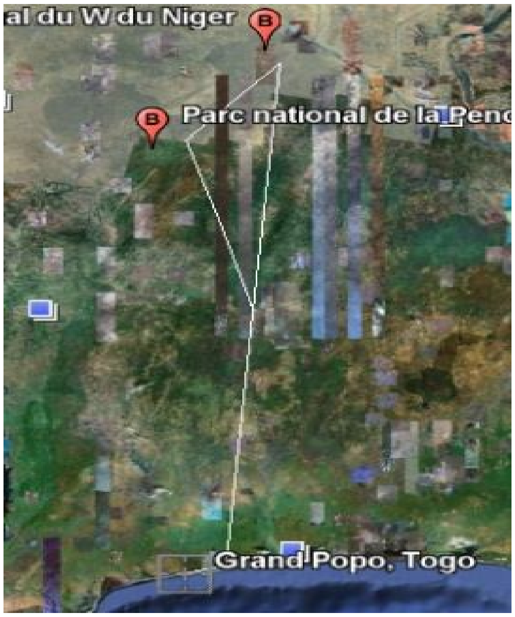


Fig. 6. Travel diary example.

4. CONCLUSIONS AND FUTURE WORK

Extracting geospatial data from texts is becoming a pressing need considering the data requirements posed by emerging Web applications utilizing geospatial data. Not wanting to rely on professional data creators, because of financial, data coverage, accuracy, etc. reasons, we will have to define tools that will allow anybody to contribute to a global geospatial data stash. This work contributes an information extraction system that (i) extracts routes from texts and (ii) goes beyond simple geocoding by actually annotating texts with routes. A main advantage of our system is that we provide plain narrative texts and we do not rely on exhaustive gazetteer lists, but a relatively small in size ontology to annotate texts and extract geospatial data. The approach is based on natural language processing techniques that provide robustness and also accuracy. Our system extracts not only route information but actual contexts of spatial objects as identified in texts. The experiments show that the proposed approach is suitable for extracting with considerably accuracy actual routes from narrative and, thus, creating geospatial data and increasing the value of the provided content.

Directions for future work are as follows. Although not examined in depth in this work, the context of spatial objects such as spatial (spatiotemporal) relationships (moving from X to Y) is identified in our proposed approach. Hence, the next step will be to map spatial relationships such as metric, topological and directional and their spatiotemporal equivalents to English language expressions and extract such data from texts (cf. [24]). A consequence of this approach will be the creation of a robust rule base for extraction of such relationships. The eventual goal of this work will be to derive arbitrary datasets such as maps automatically from texts. Here, one will have to deal with the uncertainty of user-contributed datasets and respective data fusions techniques.

5. REFERENCES

[1] Amitay, E., Har'EL, N., Sivan, R., and Soffer, A. 2004. Web-a-Where: Geotagging Web Content. In Proc. of SIGIR, 273-280.

[2] Black, W.J., McNaught, J., Vasilakopoulos, A., Zervanou, K., Theodoulidis, B. and Rinaldi, F. 2005. CAFETIERE: Conceptual Annotations for Facts, Events, Terms, Individual Entities and Relations. *Technical Report TR-U4.3.1*, January 11, University of Manchester.

[3] Borges, K. A. V. , Laender, A. H. F., Medeiros, C.B., and Davis, C.A. 2003. The Web as a Data Source for Spatial Databases. In *Proc. 4th ACM Workshop on Geographical information retrieval*, 31-36.

[4] Caster project. Open Source data binding framework for Java. http://www.castor.org. Project page.

[5] Cowie, J. and Wilks, Y. 2000. Information Extraction. In: R. Dale, H. Moisl and H. Somers (eds.) *Handbook of Natural Language Processing*, Marcel Dekker, New York.

[6] Cunningham, H., Maynard, D., Bontcheva, K., and Tablan, V. 2002. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. In *Proc. 40th Anniversary Meeting of the Association for Computational Linguistics* (ACL'02).

[7] Davidov, D. and Rappoport, A. 2009. Geo-mining: discovery of road and transport networks using directional patterns. In *Proc. 2009 Conference on Empirical Methods in Natural Language Processing*, Vol. 1, 267-275.

[8] Didion, J. Java WordNet Library (JWNL). http://sourceforge.net/projects/jwordnet. Sourceforge.net project page.

[9] Ding, J., Gravano, L., and Shivakumar, N. 2000. Computing Geographical Scopes of Web Resources. *In Proc. 26th VLDB conference*, 545-556.

[10] GeoGoogle. Google Geocoder Java API. http://geo-google.sourceforge.net/. Sourceforge project page.

[11] Google Inc. Google Maps API. http://code.google.com/apis/maps/. Web page.

[12] Hassan, A., Jones, R., and Diaz, F. 2009. A case study of using geographic cues to predict query news intent. In *Proc. 17th ACM GIS conference*, 33-41.

[13] Lieberman, M.D., Samet, H., and Sankaranarayanan, J. 2010. Geotagging with local lexicons to build indexes for textually-specified spatial data. In *Proc. 26th ICDE conference*, 201-212.

[14] Marcus, M.P., Santorini, B., and Marcinkiewicz, M. A. 1993. Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2): 313-330.

[15] MetaCarta Inc. Company homepage. http://www.metacarta.com/, Web page.

[16] McCurley, K. 2001. Geospatial mapping and navigation of the web. In *Proc. 10th WWW conf.*, 221-229.

[17] Yahoo Inc. Yahoo Yellow Pages. http://yp.yahoo.com/. Web page.

[18] Noaki, K. and Arikawa, M. 2005. A Geocoding method for natural route descriptions using sidewalk network databases," In *Proc. Web and Wireless Geographical Information Systems*, 38-50.

[19] OpenGIS® KML Encoding Standard (OGC KML). http://www.opengeospatial.org/standards/kml/. Web page.

[20] Protégé project homepage: http://protege.stanford.edu. Web page.

[21] Rinaldi, F., Dowdall, J., Hess, M., Ellman, J., Zarri, G. P., Persidis, A., Bernard, L., and Karanikas, H. 2003. Multilayer annotations in Parmenides. In *Proc. KCAP Workshop on Knowledge Mark up and Semantic Annotation*.

[22] Teitler, B. E., Lieberman, M.D., Panozzo, D., Sankaranarayanan, J., Samet, H., and Sperling, J. 2008. NewsStand: a new view on news. In *Proc. 16th ACM GIS conference*, 144-153.

[23] Waters, R. 2008. Way to go? Mapping looks to be the Web's next big thing. *Financial Times*, May 22, 2008.

[24] Xu, J. and Mark, D.M. 2007. Natural Language Understanding of Spatial Relations Between Linear Geographic Objects. *Spatial Cognition & Computation: An Interdisciplinary Journal*, 7(4):311-347.

[25] Zervanou, K. 2007. TOWL Deliverable 5.1 – Design of text feature extraction, version 1.0, October 2007. TOWL: Time-determined ontology based information system for real time stock market analysis, Technical University of Crete.

[26] Zhang, X., Mitra, P., Xu, S., Jaiswal, A.R. and Maceachren, A. 2009. Extracting Route Directions from Web Pages. In *Proc. Int'l Workshop on Web and Databases* (WebDB).

Assessment of Error in Air Quality Models Using Dynamic Time Warping

Jessica Lin
Department of
Computer Science
George Mason University
Fairfax, VA
jessica@cs.gmu.edu

Guido Cervone
Department of Geography and
Geoinformation Science
George Mason University
Fairfax, VA
gcervone@gmu.edu

Pasquale Franzese
Center for Earth Observing
and Space Research
George Mason University
Fairfax, VA
pfranzes@gmu.edu

ABSTRACT

An estimate of the error between the mean concentration of a released pollutant simulated by an atmospheric dispersion model and the values measured at the ground is obtained using Dynamic Time Warping (DTW). The error measure is relevant to the application with iterative source detection algorithms based on forward numerical transport and dispersion simulations. The new proposed measure is compared with two established error functions commonly used in the literature.

A sensitivity study of the error measure to wind direction was performed using real world data from the Prairie Grass field experiment. Whereas both standard measures found smallest error only with a few degrees of wind direction, DTW found the smallest error with a much larger range of wind directions, often as high as 20 degrees.

Categories and Subject Descriptors

H.2 [**Database Management**]: Database Applications—*Spatial databases and GIS, Data mining, Scientific databases*

General Terms

Algorithms

Keywords

Dynamic Time Warping; Time Series Analysis; Source Detection, Error Functions

1. INTRODUCTION

Detecting the source of a pollutant release in the atmosphere, and identifying its characteristics, is an important problem due to the necessity to locate the source in order to take action or to correctly assess the potential damages caused by the release. The problem can be summarized as follows. Given a few measurements of pollutant concentrations and some basic meteorological information, the goal is to identify the characteristics of the release such as location, emission mass rate, temporal evolution, in order to be able to predict the fate of the contaminants [12, 5].

Source detection algorithms can be based on backward or forward simulation techniques. Backward techniques use reverse transport and dispersion simulations from the receptor to the source. Forward techniques use transport and dispersion simulations from different candidate sources, and compare the resulting concentrations to the available measurements. The algorithms search the characteristics of the source that minimizes the error between simulated and measured concentrations. An appealing characteristic of the forward techniques is that they do not require modifications to the dispersion model. Therefore, they can be used with any available dispersion model, independent of the complexity of the problem. We will apply a forward simulation technique.

Several forward iterative methods for source estimation have been developed [9, 14, 17, 8, 5]. In particular, evolutionary or genetic algorithms were employed to drive a search process based on forward numerical simulations, and it was shown that source characteristics were correctly identified for synthetic cases and for a controlled field experiment [4, 6].

Different measures of the error between the simulated and observed values were investigated to quantify the performance of the new candidate solutions. The error function is the only feedback that the algorithm receives on the quality of the newly generated solutions. It is usually referred to as error or fitness function, and its value is also called the skill score.

The correct wind direction is paramount to source estimation problems. It was observed that errors in wind direction of only a few degrees drastically worsen the source estimation. Even when the wind direction is carefully measured at the time of the release, as for example in a field experiment, the wind variability over the time of the release can be very large leading to large uncertainty and noise in the data.

To address this problem, previous research investigated two different approaches. The first method consisted in choosing an error function that compares the simulated and observed values without taking into account their spatial distribution. In general the method performed poorly because the spatial location of the concentration plays a crucial role in correctly identifying the characteristics of the source. A second approach consisted in making the wind direction an unknown in the source estimation problem. This method generated good results, at the cost of increasing the com-

ACM SIGSPATIAL DMG Workshop, 2 Nov 2010, San Jose, CA, USA
Copyright 2010 ACM 78-1-4503-0430-6/10/11 ...$10.00.

plexity of the search process.

This paper introduces a third approach, namely the use of Dynamic Time Warping (DTW) [18, 28] to compute the error between simulated and observed concentrations. DTW is a distance measure that is commonly used in time series databases and mining [19, 26, 27, 28, 29] and signal processing communities [30, 32, 33, 31]. DTW uses dynamic programming techniques to determine the best alignment that minimizes the distance/cost/error between sequences. Its ability to produce nonlinear alignment between sequences makes it shift-invariant, and addresses the problem of errors in wind direction.

Since its introduction by Bellman in 1959 [18], DTW has been used extensively in the speech processing community [30, 32, 33, 31]. In 1994, Berndt and Clifford introduced DTW as a time series similarity measure to the database community [19, 28]. Due to its ability to minimize the effects of shifting on the time axis, DTW has been widely used in diverse fields. For example, Kuzmanic and Zanchi used DTW for hand shape (sign language) classification [20]; Corradini used DTW to recognize gestures and human activities [21]; Keogh et al. adapted DTW for various time series data mining tasks such as classification, clustering, and similarity search, on various applications such as motion capture matching and shape matching [34, 28, 29]; Niennattrakul and Ratanamahatana adapted DTW for k-means clustering for multimedia time series data [22]; Muller et al. proposed a multiscale DTW for music synchronization [23]; Aach and Church applied DTW on RNA and protein expression data [24]; and Zhang et al. compared DTW to other similarity measures for surveillance trajectory clustering [25]. While DTW is a robust similarity measure that outperforms many existing approaches, it is also computationally intensive. To mitigate this issue, several techniques for indexing DTW have been proposed [26, 27, 28]. In fact, Ratanamahatana and Keogh show that with indexing, DTW can be achieved in linear time when searching large databases [29].

This paper is structured as following: Section 2 discusses the methodology, including the different error measures used and the numerical simulation performed; Section 3 describes the experiments performed and their results; Section 4 discusses the findings and suggests applications for the proposed method.

2. METHODOLOGY

2.1 Transport and Dispersion Simulations

The dispersion simulations are performed using a Gaussian reflected dispersion model, which determines the predicted mean concentration c_s at a location x, y and z of an atmospheric tracer released from a source located at x_s, y_s, and z_s:

$$c_s = \frac{Q g_y g_z}{2\pi U[(\sigma_s^2+\sigma_y^2)(\sigma_s^2+\sigma_z^2)]^{1/2}} \tag{1}$$

with

$$g_y = \exp\left[-\frac{(y-y_s)^2}{2(\sigma_s^2+\sigma_y^2)}\right]; \tag{2}$$

$$g_z = \exp\left[-\frac{(z-z_s)^2}{2(\sigma_s^2+\sigma_z^2)}\right] + \exp\left[-\frac{(z+z_s)^2}{2(\sigma_s^2+\sigma_z^2)}\right] \tag{3}$$

where Q is the source mass emission rate, U is the wind speed, $\sigma_y(x, x_s; \psi)$ and $\sigma_z(x, x_s; \psi)$ are the crosswind and vertical dispersion coefficients (i.e. the plume spreads) where ψ describes the atmospheric stability class (i.e., $\psi = A$ to $\psi = F$), and $\sigma_s^2 = \sigma_y^2(x_s, x_s, \psi) = \sigma_z^2(x_s, x_s, \psi)$ is a measure of the area of the source. The result of the simulation is the concentration field generated by the release along an arbitrary wind direction θ. The dispersion coefficients are computed from the tabulated curves of Briggs [2].

2.2 Prairie Grass Experiment

The methodology was tested on data from the Prairie Grass field experiment [3]. The experiment consisted of 68 consecutive releases of trace gas SO_2 of 10 minutes each from a single source. The mean concentration was measured at sensors positioned along arcs radially located at distances of 50 m, 100 m, 200 m, 400 m and 800 m from the source. Information on the atmospheric conditions at the time of each release is available, and each experiment could be classified according to Pasquill's atmospheric stability classes [15, 10].

2.3 Synthetic Dataset

To test the methodology, in addition of the observed Prairie Grass measurements, we have created a synthetic dataset simulating each of the 68 releases using the model described in Equation (1) along with the meteorological and release characteristics of the original Prairie Grass experiment. The simulated concentrations are recorded at the corresponding sensor locations for the original experiments. For each of the 68 releases, a study on the effect of the wind direction was performed by varying the wind angle from -20 to +20 degrees, in a 1 degree increment. Consequently, for each of the original Prairie Grass release there are 41 synthetic releases. One case simulates the Prairie Grass experiments using exactly the parameters observed at the time of the experiments, while the other 40 vary the wind direction, and keep all other parameters constant.

This synthetic dataset allows to perform sensitivity studies to wind direction, determining for each experiment what is the wind direction that generates smaller error between the simulated and observed concentration. Even in a controlled experiment, like Prairie Grass, there are discrepancies between the observed wind direction and the angle that generates smallest error. This is because there can be errors in measuring wind at the time of the experiment, and because wind direction is usually not constant. The value reported in the official experiment summary is the average of the wind direction during the entire time of the release, and might contain errors.

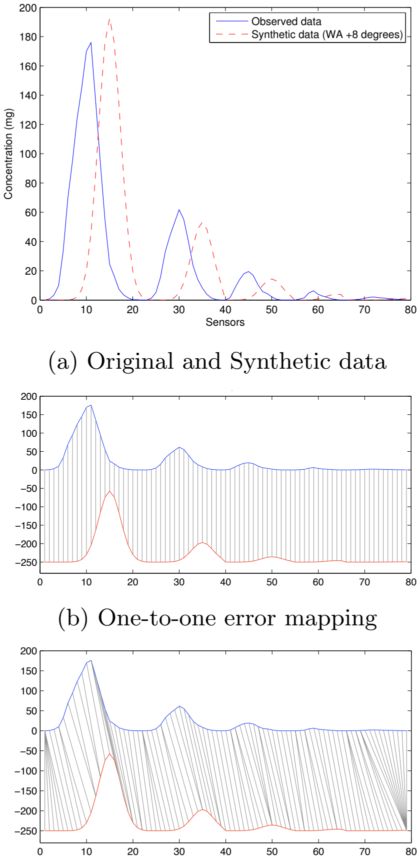
2.4 Error Functions

The quantitative comparison of observed and synthetic concentrations is performed by applying several statistical measures of error which reflect different aspects of the spatial distribution of concentration. We considered two functions: the normalized root mean square error NRMSE [11, 7], and AHY2 [13, 1, 5].

$$\text{NRMSE} = \sqrt{\frac{\overline{(c_o - c_s)^2}}{\overline{c_o}\;\overline{c_s}}} \tag{4}$$

$$\text{AHY2} = \sqrt{\frac{\overline{[\log_{10}(c_o+1) - \log_{10}(c_s+1)]^2}}{\overline{[\log_{10}(c_o+1)]^2}}} \tag{5}$$

where c_o and c_s are the observed and simulated concentration at the sensors, respectively.



(a) Original and Synthetic data

(b) One-to-one error mapping

(c) DTW error mapping

Figure 1: Example computing error between the synthetic and observed data using one-to-one and DTW mapping for Prairie Grass experiment 23. The synthetic data is shifted by eight degrees with respect to the observed parameters.

NRMSE is expressed in terms of variances, reflecting both systematic bias and relative random errors, which are estimated on a linear scale. NRMSE is strongly affected by infrequently occurring large overprediction or large observed outliers. AHY2, defined in [13] and [1] as metrics for the cost function of a genetic algorithm for source detection, computes the error on a logarithmic scale.

2.5 Dynamic Time Warping (DTW)

The error functions (4) and (5) are efficient to compute. However, both are sensitive to slight spatial distortions. To illustrate this, consider the dataset shown in Figure 1a. The observed Prairie Grass measurements and the synthetic sequence generated for the same experiment look similar in shape; however, the change in the wind direction has caused the simulated data to shift slightly to the right. This slight shifting on the x-axis will result in large errors being computed by NRMSE and AHY2, since both error functions require one-to-one mapping of data points in space. To mitigate this problem, we propose to use DTW, a well-known distance measure for signal and time series data, as our error function. Given the two sequences $X = x_1, x_2, ..., x_n$ and $Y = y_1, y_2, ..., y_m$, DTW aligns the sequences by constructing a $n \times m$ matrix M, where each entry $M(i, j)$ represents the distance $d(x_i, y_j)$ between points x_i and y_j. The entry $M(i, j)$ also corresponds to an alignment between x_i and y_j [28]. To determine the best alignment between two sequences, DTW finds a path, $W = w_1, w_2, ..., w_k$, through the matrix that minimizes the warping cost, and satisfies the following constraints [18, 28]:

1. boundary conditions: $w_1 = (1, 1), w_k = (m, n)$. This requires that the warping path starts and finishes in the first and the last points, respectively, of the sequences.

2. continuity: Let $w_i = (a, b)$ then $w_{i-1} = (a', b')$ where $a - a' \leq 1$ and $b - b' \leq 1$. This confines the allowable steps in the warping path to neighboring points.

3. monotonicity: Let $w_i = (a, b)$ then $w_{i-1} = (a', b')$ where $a - a' \geq 0$ and $b - b' \geq 0$. This requires that the points in the warping path be monotonically ordered with respect to time.

The warping cost can be computed using dynamic programming with the following recurrence [28]:

$$f(i,j) = d(x_i, y_j) + min \begin{cases} f(i-1, j-1) \\ f(i-1, j) \\ f(i, j-1) \end{cases} \quad (6)$$

In other words, the cumulative distance $f(i, j)$ is the sum of the distance between current points (x_i, y_i) and the minimum of the cumulative distance in the neighboring points [18].

Figures 1b and c show two different alignments for the dataset shown in Figure 1a. In Figure 1b, the sequences are aligned using an error function such as NRMSE that requires one-to-one mapping, i.e. no warping allowed. As the figures illustrate, the peaks are not aligned properly, thus resulting in a large error. In contrast, in Figure 1c, the sequences are aligned using DTW. The non-linear mapping allows the peaks to match, thus minimizing the error and the effect of wind direction.

Note that the concentration series that we analyze are not true time series, since the releases are continuous and the concentration field is stationary. Instead, the series describes the evolution of the concentration in space (not in time). The shift which is identified by DTW is the displacement of the simulated concentration field with respect to the observed values. Essentially, we are replacing the variable 'time' with the variable 'space', and instead of analyzing time-series, we are analyzing 'space-series'. (In this case the technique could be more correctly referred to as 'Dynamic Space Warping').

Figure 2 shows the DTW distance matrix for input sequences X and Y. Each cell (i, j) represents the distance between X_i and Y_j. The white curve along the diagonal denotes the best warping path, i.e. one that minimizes the cumulative distance.

Some global constraint on the warping path is typically specified to restrict the warping paths. The advantages of using a global constraint are two-folds: (1) it produces more intuitive alignment, and (2) it speeds up the computation by narrowing the search space. A large warping window causes the search to become prohibitively expensive, as well

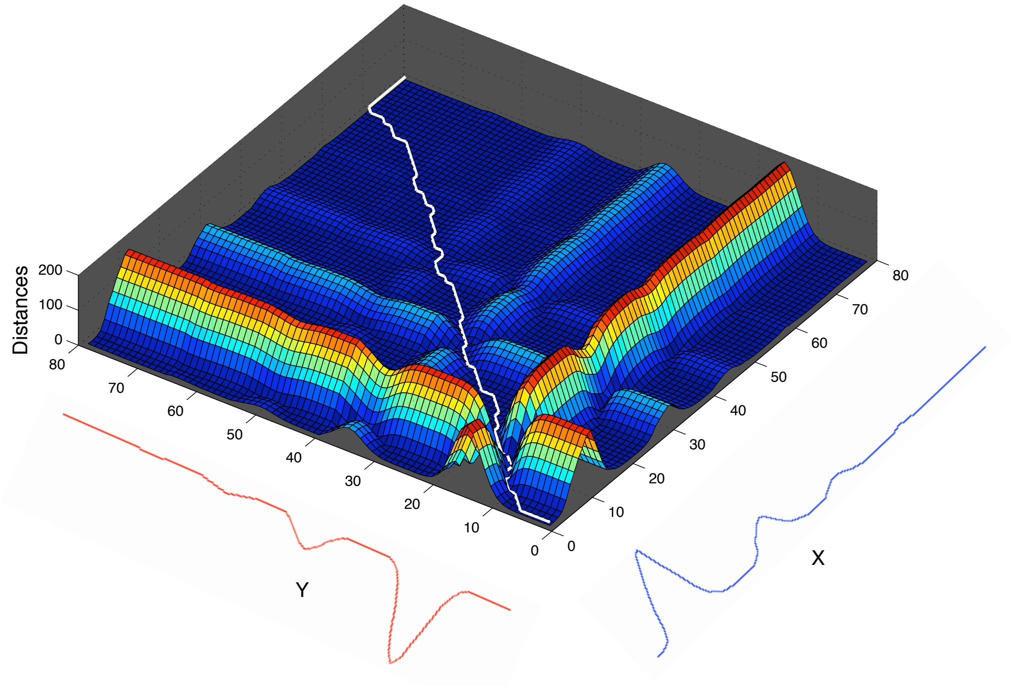


Figure 2: DTW distance matrix for input sequences X and Y. Each cell (i,j) represents the distance between X_i and Y_j. The white curve along the diagonal denotes the best warping path, i.e. one that minimizes the cumulative distance.

as possibly allowing meaningless matching between points that are far apart. On the other hand, a small window might prevent us from finding the best solution. It has been shown by Ratanamahatana and Keogh [29] that by learning the best size and shape of the global constraint for different datasets, higher accuracy can be achieved. In our work, we use the Sakoe-Chiba Band [30], with 10% of the series length as the warping window length.

3. RESULTS

The proposed method was tested by performing a sensitivity analysis of the wind angle, computing the error between the observed and simulated concentrations. Experiments were performed for all 68 Prairie Grass experiments, and repeated for each of the three methods, DTW, NRMSE and AHY2. The hypothesis is that the DTW method, because of its ability to detect shifts in time-series, is less prone to errors in wind angle. In the Prairie Grass experiments the sensors are positioned along five concentric arcs, located at 50 m, 100 m, 200 m, 400 m and 800 m from the source. The measurements are transformed into a time-series by sorting each arc counterclockwise, from the inner arc to the outer arc. The footprint of each experiment changes due to the atmospheric class, the wind characteristics and the amount of release. Therefore each release is measured by a varying number of sensors, leading to time-series that vary in length, from tenths to hundreds of observations.

Figure 3 shows the results for Prairie Grass experiment 23 (left) and 29 (right). The top figures show an interpolation of the release as measured at the time of the experiment. The sensor location are indicated with black dots, and the source of the release is at 0,0. The center panels show the results of the wind angle sensitivity analysis for each of the three methods performed. The graphs show the error between the simulated and observed concentrations as the simulated wind angle changed from -20 to $+20$ degrees with respect to the wind direction recorded at the time of the experiment. In both cases DTW obtained smaller error for a wide range of wind angles, whereas both NRMSE and AHY2 obtained best results for a very small range of values. For case 23, the best results are found when the wind angle change is close to 0, meaning that the observed concentrations are consistent with the observed wind angle. For case 29, there is a variation between 6 and 9 degrees between the observed concentrations are consistent with the observed wind angle, indicating some noise in the observed data. The bottom panels illustrate the Cross-wind profiles of concentration for the releases, plotted as functions of the sensor number. The top graph shows the observed data and the synthetic data generated using the observed wind angle. The bottom graph shows the observed data and the synthetic data generated using an adjusted wind angle. It is evident that for case 29, a wind angle of -10 degrees with respect to the observed wind direction better approximates the observed values. This is an indication of noise in the observed data, most likely to be attributed to fluctuations in wind direction during the time of the experiment.

Table 1 summarizes the results for all the 68 Prairie Grass experiments sorted by atmospheric type. The original experiment identifier (PG ID) ia also reported in the table. NRMSE and AHY2 behave similarly, finding best results with a very small (usually 1 or 2) degrees of wind angle, DTW is able to find best results with a much higher number of wind angles.

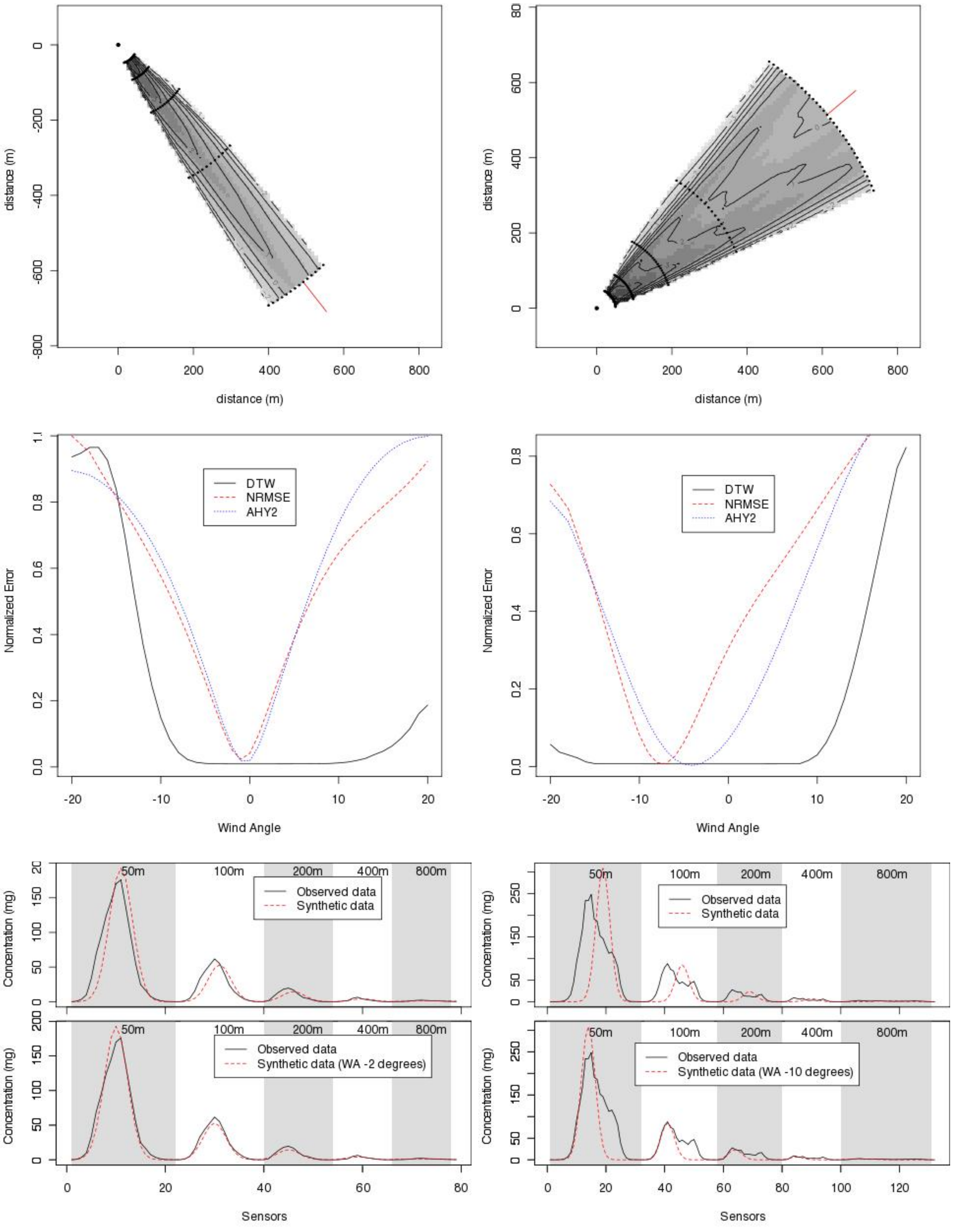


Figure 3: Results for Prairie Grass release 23 (left) and 29 (right). The graphs show a-TOP) the original releases, interpolated from the measurements made at the receptors (black circles); b-MIDDLE) the error found by AHY2, NRMSE and DTW as a function of wind angle; and c-BOTTOM) the observed and simulated concentrations using the observed wind angle, and the wind angle found by AHY2.

ψ	PG ID	DTW	NRMSE	AHY2
A	15	25	5	4
A	16	14	3	3
A	25	1	4	5
A	47	15	3	3
A	52	4	4	4
B	1	7	1	4
B	2	1	1	1
B	7	12	2	3
B	10	15	2	4
B	48S	3	2	3
C	5	19	2	3
C	8	29	2	2
C	9	23	1	3
C	19	25	2	3
C	27	27	3	3
C	43	21	3	4
C	44	13	3	4
C	49	23	2	3
C	50	23	2	2
C	62	27	2	1
D	6	29	2	2
D	11	21	1	1
D	12	27	2	2
D	17	13	3	2
D	20	29	2	3
D	21	21	2	1
D	22	17	2	2
D	23	19	1	2
D	24	23	2	1
D	26	17	3	4
D	29	25	2	3
D	30	19	2	3
D	31	21	3	3
D	33	28	2	2
D	34	25	1	2
D	35S	17	2	1
D	37	17	2	1
D	38	17	2	1
D	42	17	2	2
D	45	25	3	2
D	46	23	1	2
D	48	23	2	2
D	51	21	2	3
D	54	13	2	1
D	55	21	1	2
D	56	17	1	1
D	57	29	2	3
D	60	19	2	2
D	61	17	2	3
D	65	17	4	1
D	67	17	2	1
E	18	2	2	1
E	28	1	1	2
E	41	1	3	1
E	66	1	2	1
E	68	1	2	1
F	3	28	3	3
F	4	11	3	4
F	13	21	2	4
F	14	7	2	2
F	32	1	2	1
F	35	1	7	1
F	36	1	2	2
F	39	1	1	1
F	40	33	2	2
F	53	2	2	1
F	58	1	2	1
F	59	1	2	2

Table 1: Wind angle range in degrees for which best results were obtained using DTW, NRMSE and AHY2. The results are shown for each of the Prairie Grass experiments, identified by PG ID, and are sorted by atmospheric class

4. CONCLUSIONS

This preliminary study shows that DTW can be effectively used as the error function driving algorithms for source detection. A current shortcoming of the available error functions is that they have difficulties recognizing simple spatial shifts in the simulated distribution of concentration. This results in the error functions reporting large errors even though the simulated cloud is in fact very close to the measured one in terms of extension, shape, and magnitude, but not in the alignment. The results of the sensitivity study support the hypotheses that using DTW to compute the error between observations with simulations is less sensitive to wind direction changes. Furthermore, because some of the Prairie Grass experiments contained errors in the observed wind direction, the DTW method also works well in the presence of noise. The advantage of DTW over NRMSE and AHY2 is largest for atmospheric class A (unstable) through D (neutral). For stable atmosphere (E and F) DTW also finds best results for a rather small number of wind angles. This is most likely due to the more limited data available (smaller time-series) caused by a smaller footprint of the release, thus measured by fewer sensors.

5. ACKNOWLEDGMENTS

This material is partly based upon work supported by the National Science Foundation under Grant AGS 0849191 and by George Mason University Summer Research Funding.

6. REFERENCES

[1] C. T. Allen, G. S. Young, and S. E. Haupt. Improving pollutant source characterization by better estimating wind direction with a genetic algorithm. *Atmospheric Environment*, 41(11):2283–2289, 2007.

[2] P. S. Arya. *Air pollution meteorology and dispersion.* Oxford University Press, 1999.

[3] M. Barad. Project Prairie Grass, a field program in diffusion. Technical Report Geophysical Research Paper, No. 59, Report AFCRC-TR-58-235, Air Force Cambridge Research Center, 218pp, 1958.

[4] G. Cervone and P. Franzese. Machine learning for the source detection of atmospheric emissions. In *Proceedings of the 8th Conference on Artificial Intelligence Applications to Environmental Science*, number J1.7, January 2010.

[5] G. Cervone and P. Franzese. Monte Carlo source detection of atmospheric emissions and error functions analysis. *Computers & Geosciences*, 36(7):902–909, 2010.

[6] G. Cervone, P. Franzese, and A. Gradjeanu. Characterization of atmospheric contaminant sources using adaptive evolutionary algorithms. *Atmospheric Environment*, 44:3787–3796, 2010.

[7] J. C. Chang, P. Franzese, K. Chayantrakom, and S. R. Hanna. Evaluations of CALPUFF, HPAC, and VLSTRACK with two mesoscale field datasets. *Journal of Applied Meteorology*, 42(4):453–466, 2003.

[8] L. Delle Monache, J. Lundquist, B. Kosović, G. Johannesson, K. Dyer, R. Aines, F. Chow, R. Belles, W. Hanley, S. Larsen, G. Loosmore, J. Nitao, G. Sugiyama, and P. Vogt. Bayesian inference and Markov Chain Monte Carlo sampling to

reconstruct a contaminant source on a continental scale. *Journal of Applied Meteorology and Climatology*, 47:2600–2613, 2008.

[9] A. Gelman, J. Carlin, H. Stern, and D. Rubin. *Bayesian Data Analysis*. Chapman & Hall/CRC, 2003. 668 pp.

[10] S. R. Hanna, J. C. Chang, and G. D. Strimaitis. Uncertainties in source emission rate estimates using dispersion models. *Atmospheric Environment*, 24A(12):2971–2980, 1990.

[11] S. R. Hanna, J. C. Chang, and G. D. Strimaitis. Hazardous gas model evaluation with field observations. *Atmospheric Environment*, 27A:2265–2285, 1993.

[12] S. E. Haupt. A demonstration of coupled receptor/dispersion modeling with a genetic algorithm. *Atmospheric Environment*, 39(37):7181–7189, Dec. 2005.

[13] S. E. Haupt, G. S. Young, and C. T. Allen. Validation of a receptor/ dispersion model coupled with a genetic algorithm using synthetic data. *Journal of Applied Meteorology*, 45:476–490, 2006.

[14] G. Johannesson, B. Hanley, and J. Nitao. Dynamic bayesian models via monte carlo - an introduction with examples -. Technical Report UCRL-TR-207173, Lawrence Livermore National Laboratory, October 2004.

[15] F. Pasquill and F. Smith. *Atmospheric Diffusion*. Ellis Horwood, 1983.

[16] S. K. Rao. Source estimation methods for atmospheric dispersion. *Atmospheric Environment*, 41(33):6964–6973, 2007.

[17] I. Senocak, N. Hengartner, M. Short, and W. Daniel. Stochastic event reconstruction of atmospheric contaminant dispersion using Bayesian inference. *Atmospheric Environment*, 42(33):7718–7727, 2008.

[18] R. Bellman and R. Kalaba On Adaptive Control Processes. *IRE Transactions on Automatic Control*, Vol. 4, No. 2, pp.1 9, 1959.

[19] D. Berndt and J. Clifford Using Dynamic Time Warping to Find Patterns in Time Series. In *Proceedings of the AAAI-94 Workshop on Knowledge Discovery in Databases*, pp. 229-248.

[20] A. Kuzmanic and V. Zanchi Hand Shape Classification Using DTW and LCSS as Similarity Measures for Vision-Based Gesture Recognition System. In *Proceedings of EUROCON, 2007, the International Conference on Computer as a Tool*, pp. 264-269.

[21] A. Corradini. Dynamic Time Warping for Offline Recognition of a Small Gesture Vocabulary. In *Proceedings of the IEEE ICCV Workshop on Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems*, July 13. Vancouver, Canada.

[22] V. Niennattrakul and C. A. Ratanamahatana. On Clustering Multimedia Time Series Data Using k-Means and Dynamic Time Warping. In *Proceedings of the 2007 International Conference on Multimedia and Ubiquitous Engineering*, pp. 733-738. Apr 26-28, Seoul, Korea.

[23] M. Muller, H. Mattes and F. Kurth. An Efficient Multiscale Approach to Audio Synchronization. In *Proceedings of the 7th International Conference on Music Information Retrieval (ISMIR)*. pp. 192-197. Oct 10-12, 2006. Victoria, Canada.

[24] J. Aach and G. Church. Aligning Gene Expression Time Series with Time Warping Algorithms. *Bioinformatics*, 17:495-508.

[25] Z. Zhang, K. Huang and T. Tan. Comparison of Similarity Measures for Trajectory Clustering in Outdoor Surveillance Scenes. In *Proceedings of the 18th International Conference on Pattern Recognition (ICPR'06)*. pp. 1135-1138. Washington DC.

[26] B. Yi, K. Jagadish and C. Faloutsos. Efficient Retrieval of Similar Time Sequences Under Time Warping. In *Proceedings of the 14th International Conference on Data Engineering (ICDE)*. Feb 23-27, 1998. Orlando, FL.

[27] S. Kim, S. Park and W. Chu. An Index-Based Approach for Similarity Search Supporting Time Warping in Large Sequence Databases. In *Proceedings of the 17th International Conference on Data Engineering (ICDE)*. Apr 2-6, 2001. Heidelberg, Germany.

[28] E. Keogh. Exact Indexing of Dynamic Time Warping. In *Proceedings of the 28th International Conference on Very Large Data Bases (VLDB)*. Hong Kong, China. August 20-23, 2002.

[29] C. A. Ratanamahatana and E. Keogh. Three Myths About Dynamic Time Warping. In *Proceedings of SIAM International Conference on Data Mining (SDM'05)*. pp. 506-510. Newport Beach, CA. April 21-23.

[30] H. Sakoe and S. Chiba. Dynamic Programming Algorithm Optimization for Spoken Word Recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing*. 26(1):43-49.

[31] L. Rabiner and B. Juang. *Fundamentals of Speech Recognition*. Englewood Cliffs, NJ, Prentice Hall.

[32] L. Rabiner, A. Rosenberg and S. Levinson. Considerations in Dynamic Time Warping Algorithms for Discrete Word Recognition. *IEEE Trans. Acoustics, Speech, and Signal Processing*, Vol. ASSP-26, 575-582. 1978

[33] C. Myers, L. Rabiner and A. Rosenberg. Performance Tradeoffs in Dynamic Time Warping Algorithms for Isolated Word Recognition. *IEEE Trans. Acoustics, Speech, and Signal Processing*, Vol. ASSP-28, 623-635.

[34] E. Keogh and M. J. Pazzani. Scaling Up Dynamic Time Warping to Massive Datasets. In *Proceedings of the 3rd European Conference on Principles of Data Mining and Knowledge Discovery*, pp. 1-11. September 15-18, 1999. Prague, Czech Republic.

View Reconstruction from Images by Removing Vehicles

Li Chen
Department of Electrical and Computer Engineering,
Virginia Tech
4300 Wilson Blvd., Ste. 750
Arlington, VA 22203
703-387-6060
lchen06@vt.edu

Lu Jin
Department of Electrical and Computer Engineering,
Virginia Tech
4300 Wilson Blvd., Ste. 750
Arlington, VA 22203
202-687-7451
lujin@vt.edu

Jing Dai
T.J. Watson Research Center
IBM
19 Skyline Dr.
Hawthorne, NY 10532
914-784-6460
jddai@us.ibm.com

Jianhua Xuan
Department of Electrical and Computer Engineering,
Virginia Tech
4300 Wilson Blvd., Ste. 750
Arlington, VA 22203
703-387-6060
xuan@vt.edu

ABSTRACT

Reconstructing views of real-world from satellite images, surveillance videos, or street view images is now a very popular problem, due to the broad usage of image data in Geographic Information Systems and Intelligent Transportation Systems. In this paper, we propose an approach that tries to replace the differences among images that are likely to be vehicles by the counterparts that are likely to be background. This method integrates the techniques for lane detection, vehicle detection, image subtraction and weighted voting, to regenerate the "vehicle-clean" images. The proposed approach can efficiently reveal the geographic background and preserve the privacy of vehicle owners. Experiments on surveillance images from TrafficLand.com and satellite view images have been conducted to demonstrate the effectiveness of the approach.

Categories and Subject Descriptors

I.2.10 [**Artificial Intelligence**]: Vision and Scene Understanding –*Modeling and recovery of physical attributes.*

General Terms

Algorithms, Experimentation.

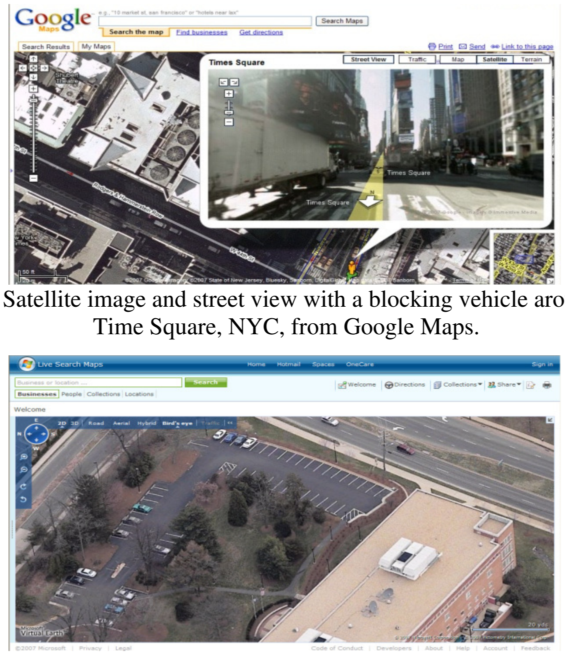
Keywords

View reconstruction, Vehicle detection, Gabor wavelet filtering.

1 . INTRODUCTION

Reconstructing view of real-world from satellite images, surveillance videos, or street view images is now a very popular problem, due to the broad usage of image data in Geographic Information Systems and Intelligent Transportation Systems. Hot map applications, including Google Maps [1] and Microsoft Live Search Maps [2], provide tremendous aerial images and street view/bird's eye images (as shown in Figure 1).

These images usually contain a lot of vehicles on/off the roads and in parking lots, which sometimes obstruct the geographic view of terrain and meaningful objects, such as lane markers, traffic signs, and fire faucets. In some cases, these vehicles may cause the leakage of private information to the public, because it is possible to refer a person's activity by looking at his/her vehicle in these images with obvious location information. On the other hand, traffic systems utilize videos or image sequences from surveillance cameras to analyze traffic and track vehicles. Reconstructing the background image helps effectively identifying the moving vehicles and automatically measuring traffic. Background learning techniques are emerged in these applications to automatically reconstruct the clean views and preserve privacy.



a). Satellite image and street view with a blocking vehicle around Time Square, NYC, from Google Maps.

b). Bird's eye image of a campus faculty parking lot with personal vehicles from Microsoft Live Search Maps.

Figure 1. Examples of online map systems.

Several effective approaches have been proposed to extract background information from video/image sequences by image subtraction/averaging. However, they usually require a large set of images to get a complete background image, which makes them not suitable for satellite images or street view images. This paper focuses on reconstructing the static background information by removing vehicles from a small set of sample images (as few as 2). We propose an approach that tries to replace the differences among images that are likely to be vehicles by the counterparts that are likely to be background. This method integrates the techniques for lane detection, vehicle detection, image subtraction and weighted voting, to regenerate the "vehicle-clean" images.

ACM SIGSPATIAL DMG Workshop, Nov 2, San Jose, CA, U.S.A.
Copyright 2010 ACM 978-1-4503-0430-6/10/11...$10.00.

The proposed approach can efficiently reveal the geographic background and preserve the privacy of vehicle owners. Experiments on surveillance images from TrafficLand.com [3] and satellite view images have been conducted to demonstrate the effectiveness of the approach.

The rest of this paper is organized as follows. Section 2 surveys the existing work on background learning methods. The proposed approach, including lane/parking lot detection, vehicle detection, image subtraction, and weighted voting, is presented in Section 3. Section 4 illustrates the experiment results. Finally, this work is concluded in Section 5.

2 . RELATED WORK

Background learning methods are widely used as the fundamental for moving objects identification. The existing methods can be divided into two categories, frame-oriented and pixel-oriented, which are both based on image sequences. The frame-oriented [4] methods take one image from the sequence as the background if its difference with the predecessor is less than a threshold. This category of background learning methods extracts the background efficiently and successfully for indoor environments. But it's not sufficient for outdoor scenarios that have noises from weather and illumination. Furthermore, in crowded urban area, it's difficult to get one picture as static background without moving objects.

The pixel-oriented approaches [5-7] collect a set of surveillance images and then either average or subtract and average the images to iteratively construct the background. These approaches usually require a large set of images, because for each pixel in the images, the background samples have to beat the vehicle samples in quantity. The requirement becomes even seriously in crowded area, where the chance to see the background is small. This also limits the extendibility of these approaches to the satellite images and street view images, because it is difficult/expensive to acquire a large number of images for the same location. On the other hand, noises from off-road objects (e.g., trees in wind) and on-road objects (e.g., changing traffic lights) could ruin or delay the results.

The proposed approach fuses the vehicle detection into background reconstruction. Therefore, the background pixels can be identified earlier to save the number of learning samples. In ideal situation, our approach requires only two image samples to generate the background. This algorithm is not only sufficient for surveillance images, but also suitable for satellite and street view images.

3 . VIEW RECONSTRUCTION ALGORITHM

This paper proposes an approach to learn the background from a couple of images. Different to the existing approaches, the proposed method considers the road region and vehicle blobs to refine the moving object identification, rather than just replying on the image subtraction/averaging. This framework makes the following assumptions.

- The input images are satellite images, surveillance images, or street view images.
- The input images have small sample size, and are not necessarily to be sequential.
- The input images have already been normalized and calibrated, so that all the samples have the same illumination, view area, and view angle.

The view reconstruction framework has four major components, namely, lane/parking lot detection, vehicle detection, image subtraction, and weighted voting, as illustrated in Figure 2. The first component, lane/parking lot detection, takes a set of (at least 2) image samples as inputs, and outputs the detected road/parking lot area. Taking the road/parking lot area and the original image samples, the vehicle detection component identifies the possible image segments that may represent vehicles. Due to the different characteristics, we developed two different vehicle detection methods for street view and satellite view, respectively. Meanwhile, the image subtraction detects the differences around the road area among image samples and marks these differences as possible moving objects. The final component, weighted voting, collects the potential vehicle segments and potential moving objects and replaces the potential moving objects with non-vehicle segments. The design of applying two parallel components, vehicle detection and image subtraction is because each single component has limited ability to accurately identify moving vehicles. The vehicle detection could sometimes regard static objects as possible vehicles. On the other hand, the image subtraction may include accidentally moving objects or changing environment as false positive results. Combining these two components, the accuracy of view reconstruction can be improved. The details for each component are discussed in the following sections.

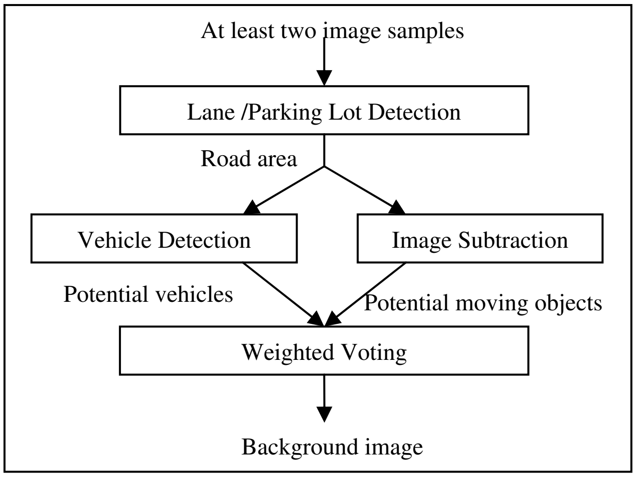


Figure 2. Background reconstruction algorithm overview.

3.1 Lane/Parking Lot Detection

The objective of lane detection is to find the lane area in a set of images which taken from the same view. The procedure of this module can be illustrated in Figure 3, which includes edge detection, line detection by Hough transform, and road detection. A similar approach can be used to detect parking lots.

3.1.1 Edge detection

Edges are often used in image analysis for finding region boundaries. Edges are defined as the pixels where the intensity changes abruptly. Many edge detectors have been used in edge detection to find the edges. In this paper, we focus on Canny edge detector. The Canny edge detection is a multi-stage algorithm to detect a wide range of edges in images. Edge detection is very useful in many image processing applications. In this work, we need it to get the edge pixels in the following Hough transform to detect lines in the image.

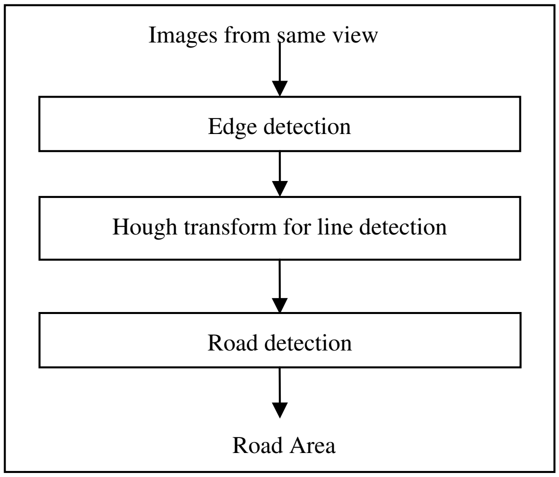


Figure 3. Flow chart of lane detection procedure.

3.1.2 Hough transform

The Hough transform is a widely used feature extraction technique to find imperfect instances of objects within a certain class of shapes by a voting procedure [8]. This voting procedure is carried out in a parameter space, from which object candidates are obtained as local maxima in a so-called accumulator space that is explicitly constructed by the algorithm for computing the Hough transform. Hough transform has the advantage of solving the problem of missing/noisy points or pixels on the desired curves (e.g. lines, circles or ellipses) due to imperfections in either the image data or the edge detector. The idea of Hough transform is to group edge points into object candidates by performing an explicit voting procedure over a set of parameterized image objects.

3.1.3 Road detection

Since we have several satellite images obtained from same view, we implement lane detection based on multiple images in order to reduce the false lines. Only those parameters among the candidate local maximum parameters are determined as the lines where the parameters have been detected in the majority of images. After determine the lines, we implement a scan approach to determine the road area. The basic idea is to decide the pixel as road area when the distance between its two neighbored lines is greater than some threshold. In the algorithm, we scan the pixels in a vertical and horizontal way, respectively to find the lane area, which could be implemented by following procedure:

***Algorithm 1**. Road Detection Procedure*

```
1. horizontal scan
   For each line $i$,
     Calculate the distance between each two adjacent line pixels along the line.
       For each pixel $(i, j)$
           Determine the distance of its neighbored lines $d_i$
         If $d_i > \Delta d$
             $(i, j)$ is a road pixel
         Else
             $(i, j)$ is not a road pixel
2. vertical scan
   For each line $j$,
     Calculate the distance between each two adjacent line pixels along the line.
       For each pixel $(i, j)$
           Determine the distance of its neighbored lines $d_j$
         If $d_j > \Delta d$
             $(i, j)$ is a road pixel
         Else
             $(i, j)$ is not a road pixel
```

This approach has the problem that when the lane appears at the corner because some parts of lane will be misclassified (illustrated in Figure 4). So we improved this method by adding the boundary of the images for those corners which may have lanes. If $\theta < 90^o$, the left bottom and right up corners have the possibility to have the lanes so that the boundaries need to be added. On the other hand, if $\theta > 90^o$, the other two corners may have the lanes and the boundaries need to be added.

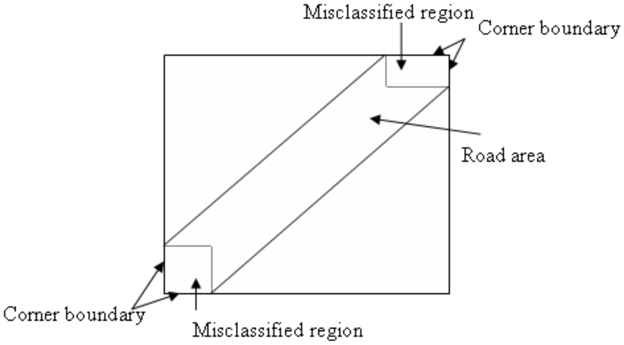


Figure 4. Illustration of improved scanning algorithm for road detection.

3.1.4 Parking Lot Detection

For most of the parking lot area, its boundary is marked by white color exactly like the road area. So we could use the same approach to acquire the parking lot area. The only different is that, a certain parking lot must have multiple entrances and exits, which may break the boundary line for a short distance, in order to avoid that we could slightly enlarge the value of Δd to acquire the correct parking lot area.

3.2 Vehicle Detection for Street View

The vehicle detection component is to find the blobs in the images that may represent vehicles or parts of vehicles. To achieve this target, we combine the color information and different levels of feature information of vehicle to generate all the blobs in the images. The procedure of vehicle detection for street view is illustrated in Figure 5. Specifically, based on color information and a set of training images, we first classify the image pixels into vehicle or non-vehicle pixels. Then the candidate vehicle blobs will be identified according to the detected road area and some domain knowledge. Finally, we extract the edge information and Gabor wavelet coefficient to further verify if the candidate blobs are from real vehicles using a classifier trained on training samples. The vehicle pixel detection, vehicle blob detection and verification steps will be discussed in detail in following subsections.

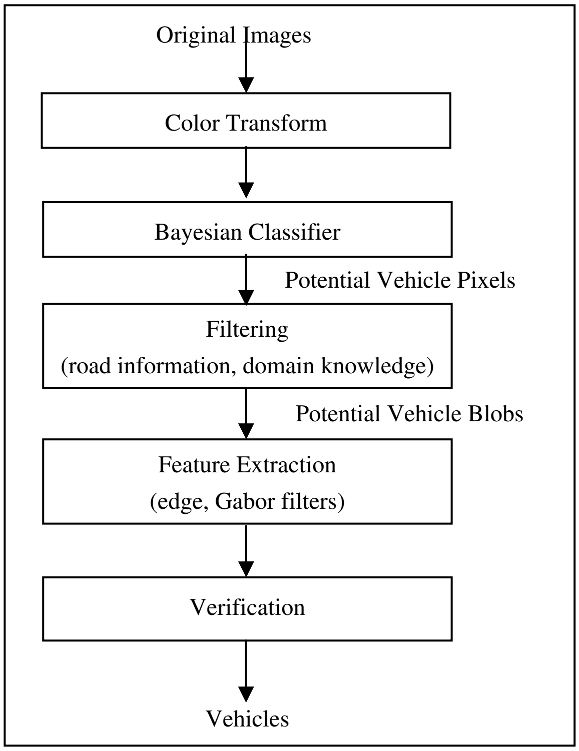


Figure 5. Flow chart of vehicle detection for street view.

3.2.1 Vehicle pixel detection

First, training images are collected from different senses including roads, parking lots, etc., as vehicles and non-vehicle (background) image sets. Figure 6 shows parts of the training samples. Based on the training samples, we applied a color transform method to reduce the original three color component (RGB) to two normalized components (UV) through following equations:

$$u_p = \frac{2Z_p - G_p - B_p}{Z_p} \quad (1)$$

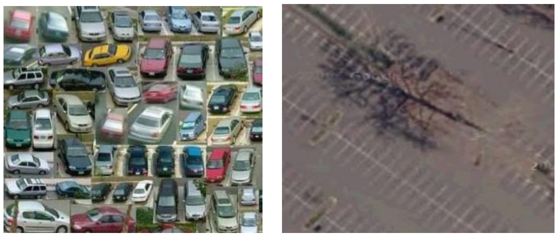
$$v_p = \max\left\{\frac{B_p - G_p}{Z_p}, \frac{R_p - G_p}{Z_p}\right\} \quad (2)$$

where (R_p, G_p, B_p) is the color of a pixel p and $Z_p = (R_p + G_p + B_p)/3$ used for normalization. Paper [9] has demonstrated that all the road colors will concentrate around a small circle through this transformation, as shown in Figure 7.

After getting the transformation color components, we then build a Bayesian classifier to identify vehicle pixels from backgrounds with colors. We assume that the RGB color components in the (U, V) domain forms a multivariate Gaussian distribution. The parameters of two Gaussian distributions are learned by EM algorithm. Assume that the mean colors of the vehicle and non-vehicle pixels are m_v and m_n, respectively. In addition, Σ_v and Σ_n are their corresponding covariance matrices in the same color domain, respectively. Then, given a pixel x, based on Bayesian classifier, we assign it to class 'vehicle' if $d_v(x) - d_n(x) > \lambda$, where

$$d_c(x) = (1/2)(x - m_c)\Sigma_c^{-1}(x - m_{c.})^t,\ c \in \{v, n\}$$

$$\lambda = \log[\sqrt{|\Sigma_v|/|\Sigma_n|}(P(nonveh)/P(vehicle))]$$



(a) Vehicles (b) Non-vehicles

Figure 6. Part of training images.

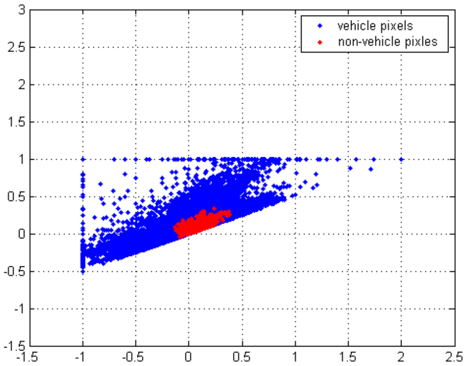


Figure 7. Color transformations of vehicle pixels and non-vehicle pixels on training images.

Finally, we can get a number of segments on the original image where the majority of pixels in the segments are classified as vehicle pixels.

3.2.2 Vehicle blob detection

The vehicle blob detection takes the domain knowledge and the road information to define a set of rules to filter out the segments that are not likely to be part of vehicle. From the road information extracted in the lane detection component, we know the area of the road, as well as the range of lane width on the image. Then from the domain knowledge, we may get the idea about the location of vehicles, their range of size and length regarding to the lane width. In our implementation, we set the following rules to filter segments:

- The possible vehicle segments have to touch the road area.
- The width (size on x-axis) of a possible vehicle segment cannot be greater than 1/3 of the road width.
- The length (size on y-axis) of a possible vehicle segment cannot be greater than 1/3 of the road width.
- The total size (number of pixels) of a possible vehicle segment cannot be greater than 1/6×1/6 of the road width, or smaller than 10.
- The intensity of a possible vehicle segment cannot be similar to the largest segment in the image (supposed to be background).

Using the above rules, we can filter out the majority of the segments that are not expected to represent parts of the vehicles. The remaining segments will be used in the final weighted voting to identify the pixels should be removed.

3.2.3 Vehicle verification

After obtaining potential vehicle blobs, we further verify if these blobs are real vehicles based on their features. The extracted features include edge and coefficients of multi-channel and multi-resolution Gabor filtering. We assume that the vehicle blobs should have rich and orientated edge information. The multi-channel approach in Gabor filtering uses a bank of band-pass filters to decompose an original image into several filtered images which contain information in different orientations and frequency ranges. This set of filtered images provides the orientation sensitive information which is essential for vehicle identification. The multi-resolution approach, on the other hand, provides local texture features at various resolutions and reliable results.

A Gabor filter can be viewed as a sinusoidal plane of particular frequency and orientation, modulated by a Gaussian envelope [4]. It can be written as:

$$h(x, y) = s(x, y)g(x, y) = h_R(x, y) + jh_I(x, y) \quad (3)$$

where $s(x,y)$ is a complex sinusoid, known as a carrier, and $g(x,y)$ is a 2-D Gaussian shaped function, known as envelope. The complex sinusoid is defined as follows,

$$s(x, y) = \exp(-j2\pi(u_0 x + v_0 y)) . \quad (4)$$

The 2-D Gaussian function is defined as follows,

$$g(x, y) = \frac{1}{2\pi\sigma\beta}\exp(-\frac{1}{2}(\frac{x^2}{\sigma^2} + \frac{y^2}{\beta^2})) . \quad (5)$$

The Gabor filter is a bandpass filter centered on frequency (u_0, v_0), with a bandwidth determined by (σ, β). Usually, a radial frequency $f = \sqrt{u_0{}^2 + v_0{}^2}$, with orientation $\theta = \tan^{-1}(v_0 / u_0)$, are used in polar coordinates to specify the filter. The Gabor filtered output of an image I is obtained by the convolution of the image with the specified Gabor function. The local energy is defined as,

$$E = C_R^2 + C_I^2 ,$$

where

$$C_R = I \oplus h_R , \quad C_I = I \oplus h_I .$$

Note that when frequency f = 0, the Gabor filter reduces to a Gaussian filter which can capture the low pass frequency of the input image. The different orientation channels in the frequency domain could be obtained by rotating θ.

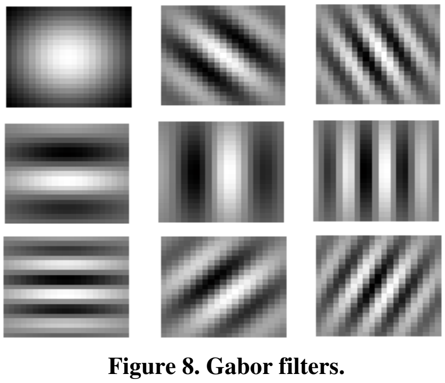


Figure 8. Gabor filters.

3.3 Vehicle Detection for Satellite View

Image segmentation algorithms generally are based on one of the two basic properties of intensity values: discontinuity and similarity. In the first criteria, the image is partitioned when there is an abrupt change in the intensities of neighboring pixels, such as edges. In the second criteria, the image is partitioned into regions which are similar according to some predefined conditions. A simple way to do image segmentation is by using thresholding techniques, and this is the technique that is used for the detection of the vehicles in satellite view. To achieve our goal of detecting all the vehicles on the satellite images, we have divided the algorithm in two parts [10, 11]:

- Detection of vehicles which are brighter than the background.
- Detection of vehicles which are darker than the background.

3.3.1 Thresholding technique

Thresholding is one of the widely methods used for image segmentation. It is useful in discriminating foreground from the background. By selecting an adequate threshold value *T*, the gray level image can be converted to binary image. The binary image should contain all of the essential information about the position and shape of the objects of interest (foreground). The advantage of obtaining first a binary image is that it reduces the complexity of the data and simplifies the process of recognition and classification. The most common way to convert a gray-level image to a binary image is to select a single threshold value (*T*). Then all the gray level values below this T will be classified as black (0), and those above *T* will be white (1). The segmentation problem becomes one of selecting the proper value for the threshold *T*. A frequent method used to select *T* is by analyzing the histograms of the type of images that want to be segmented. The ideal case is when the histogram presents only two dominant modes and a clear valley (bimodal). In this case the value of *T* is selected as the valley point between the two modes. In real applications histograms are more complex, with many peaks and not clear valleys, and it is not always easy to select the value of *T*.

In the present work our sample images contain scenes of parking lot, and our goal is to apply segmentation techniques to detect and classify the vehicles on it.

3.3.2 Detection of bright vehicles

Two different methods are proposed to identify the bright vehicles:

- Multiple Thresholding
- Clustering by Otsu Method

3.3.2.1 Computing multiple threshold

By analyzing several of the sample images in the database, we can say that the intensity values of bright vehicles are greater than the intensities of the background, and sometimes there is a region where they overlap. Because of the overlapping, some objects or regions on parking lot, such as lane markers, parking lot dividers, may have intensity values similar to the intensity of some of the bright vehicles. Also, each bright vehicle may not have same range of intensity because of atmospheric conditions and different color of the vehicles. Therefore, to identify only the vehicles and to avoid the detection of irrelevant objects like lane markers, three different thresholds *T*1, *T*2, and *T*3 are used. The algorithm is shown in below:

Algorithm 2. *Multiple thresholding algorithm*

```
M: two-dimensional matrix which stores intensity image
V: A vector calculated by M
  For each row i in M
    V[i] = maximum intensity of i-th row in M
T1 := mean of V
T2 := minimum of V
T3 := mean of T1 and T2
```

Thresholds *T*1, *T*2, and *T*3 are used to convert the test image to three different binary images Image1, Image2, and Image3. For a pixel at coordinates (x, y), if its intensity $I(x, y) > T$, then it is considered as an Object Pixel (1) else Background Pixel (0).

To avoid or reduce the detection of irrelevant objects such as lane markers and parking lot dividers, logical operations are performed among the binary images. In the case of detecting bright vehicles, best results are obtained by taking common objects from pairs of the binary images generated by the thresholds *T*1, *T*2, and *T*3. This means, common objects from the pairs: (Image1, Image2), (Image2, Image3), and (Image1, Image3) are taken. Each of these binary images may contain some or all vehicles, so add all these images together to obtain all the possible bright vehicles.

3.3.2.2 Clustering by Otsu Method

The Otsu threshold uses class separatability, and maximizes the between-class variance to find an optimal threshold value *k**. This threshold value is used to extract objects from their background. Apply the Otsu threshold to the test image directly, will detect the bright vehicles, but also some of the lane markers and parking lot dividers that are present on the parking lot. To reduce the complexity of identifying lane markers and parking lot dividers, a preprocess step is applied first. The preprocess step involves the application of a sliding neighborhood operation to the test image. The sliding neighborhood operation consists of assigning to each pixel of the test image, the maximum intensity of its neighborhood - this is a rectangular area of 3×3 pixels, with the center pixel as the one that is being processed by the operation.

By applying sliding neighborhood operation to the test images, the bright pixels corresponding to large objects, such as vehicles, become brighter, but the bright pixels corresponding to irrelevant objects such as lane markers stay at about the same level of brightness. This preprocessing step will help to highlight the vehicles, and dim some of the irrelevant objects such as the lane markers.

3.3.3 Detection of dark vehicles

For the detection of dark vehicles, the Otsu Threshold is also employed. Before applying the Otsu Threshold, a sliding neighborhood operation is applied to the test image. Because in this case we want to detect dark vehicles, each pixel is assigned with the minimum intensity of its neighboring pixel in a rectangular neighborhood of a 3×3 matrix. As a result, dark vehicles become darker when compared to the background. After applying the sliding neighborhood operation, the Otsu Threshold is used to convert the test image to a binary image.

To avoid considering vehicle's shadows as dark vehicles, the results of applying both algorithms (detection of bright vehicles and the detection of dark vehicles) are combined. This is carried out by performing a logical OR operation to the two binary images: the first one obtained by the bright vehicles detection algorithm (Otsu or Multiple Thresholds technique), and the second one obtained by the dark vehicles detection algorithm (Otsu Threshold technique). As a result of the addition, shadows of the bright vehicles, which are very close to the bright vehicles, are combined as a single vehicle. But, in cases where dark and bright vehicles are placed very close to each other, it is possible that a bright vehicle could be combined with a dark vehicle and counted as a single vehicle, giving an error in the result. This is also true when two or more bright or dark vehicles are very close to each other, then the final result would combine this group of very close vehicles as a single vehicle, causing an error in the final result.

3.4 Image Subtraction

Image subtraction identifies the differences between any pair of image samples. This is a technique that has been widely applied in the background learning approaches. The subtraction procedure compares every two pixels in the same position from two images, if the difference in intensity is greater than a threshold, marks it as "inconsistent" (notated by 0), otherwise keeps the original value from the first input image. This procedure can be formally presented as follows.

$$I'_{x,y} = \begin{cases} I^1_{x,y}, & if\ |I^1_{x,y} - I^2_{x,y}| < T; \\ 0, & if\ |I^1_{x,y} - I^2_{x,y}| >= T. \end{cases}$$

$I'_{x,y}$ indicates the intensity of the pixel at row x and column y in the result image, $I^k_{x,y}$ means the intensity of pixel at row x and column y in the k-th input image, and T is a user defined threshold. This technique usually outputs a lot of noises due to the road-side objects. We process the subtraction only around the road area to reduce the noises.

3.5 Weighted Voting

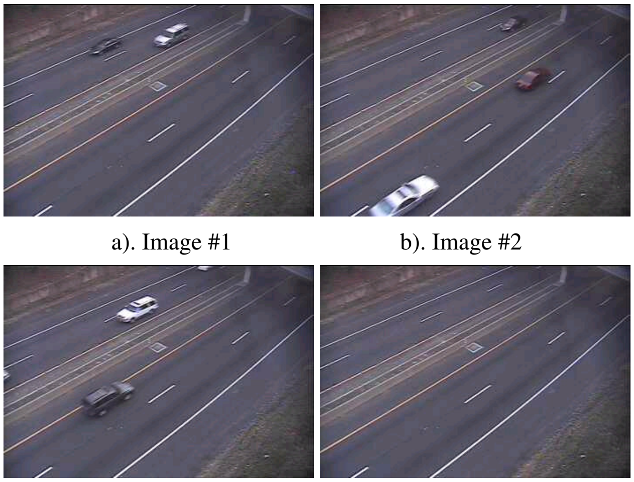
The final voting component is to replace possible moving vehicle pixels with appropriate background pixels. Traditionally, the voting technique runs a majority voting among sample images for each zero-value pixel identified by image subtraction. In this algorithm, since the results from both vehicle detection and image subtraction components need to be combined, a weighted voting strategy is developed to generate the final view. The weighted voting takes the following steps on image subtraction results and vehicle detection results.

Algorithm 3. *Weighted Voting Procedure*

```
For each zero-value pixel p identified by one image subtraction result
    For each vehicle detection results from image sample I
        If p is contained in possible-vehicle segments in I
            Weight I by multiplying d (d < 1)
    Run majority voting for p on all available samples
Return the image with replaced pixels
```

4 . EXPERIMENTS

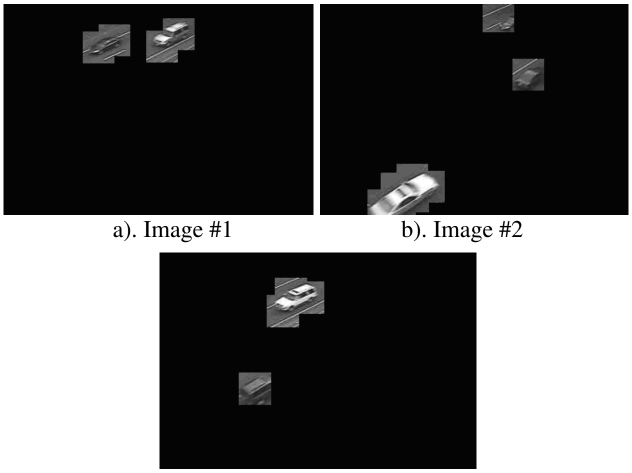
To verify the view reconstruction algorithm, surveillance images from TrafficLand.com [3] and satellite images from Googla Maps [1] have been collected. The surveillance images show the road situation of both directions on I-66 around Glebe Road with almost the same view area and illumination. Each image sample is a highly compressed image with resolution of 352×240, as shown in Figure 9. These images have low quality, but very small size for web transferring. The results of vehicle detection component are displayed in Figure 10. We can find that in our experiments, the vehicle detection procedure is very effective to detect the potential vehicle parts.



a). Image #1 b). Image #2

c). Image #3 d). Expected background

Figure 9. Original surveillance images and expected results.



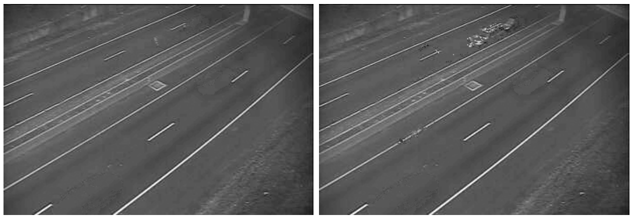
a). Image #1 b). Image #2

c). Image #3

Figure 10. Detected vehicles.

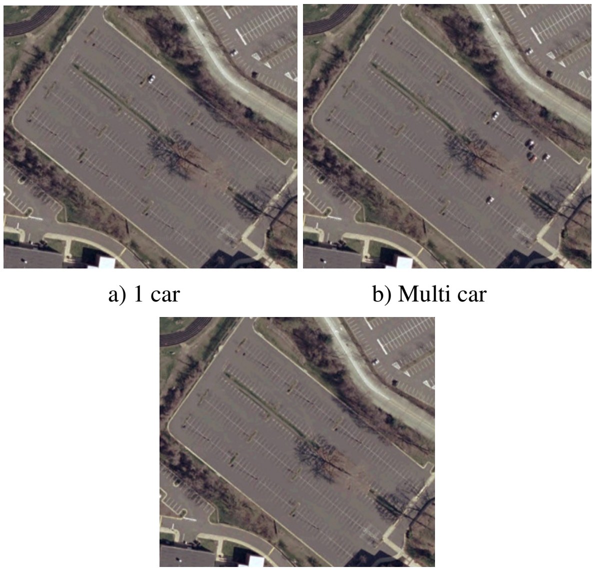
The final background image is illustrated in Figure 11, where the background generated using majority voting is displayed for comparison. In most area of the result images, the background is reconstructed well. However, around the top left corner of the road, the weighted voting approach does not remove the vehicle pixels completely, while the majority voting returns better results. We think that was because of the low accuracy of the vehicle detection results (shown in Figure 10 (b)). As the vehicles appearing around that area are usually small, and the illumination is also weak on that part, the vehicle detection approach we applied cannot identify the vehicle segments well. This problem should be relieved if image quality is improved. Further fine tuning on the image sets and the reconstruction algorithm is needed to make this approach more applicable in real-world scenarios.

For the Satellite view vehicle detection, Figure 12 shows the dataset we used for this paper, Figure 13 shows the results for detecting bright vehicles and dark vehicles using the algorithm described above. From the experiment results, we can notice that the dark vehicle is much harder to detect because the blob of the dark vehicle is smaller than the bright one, so some of the dark vehicles are hard to recognize in the binary level image and sometimes may easily be treated as noise of the image, such as the shadow of the tree. In Figure 13 the detected vehicles are pointed out with red circles. We didn't use complicated images for this experiment, so the issues we described in Section 3.3 did not happen during the process. For the reconstruction part, satellite view images could be treated the same way as street view image. We just need to replace these pixels which belong to the vehicle by background pixels.



a). Majority voting b). Weighted voting

Figure 11. Re-constructed background image.



a) 1 car b) Multi car

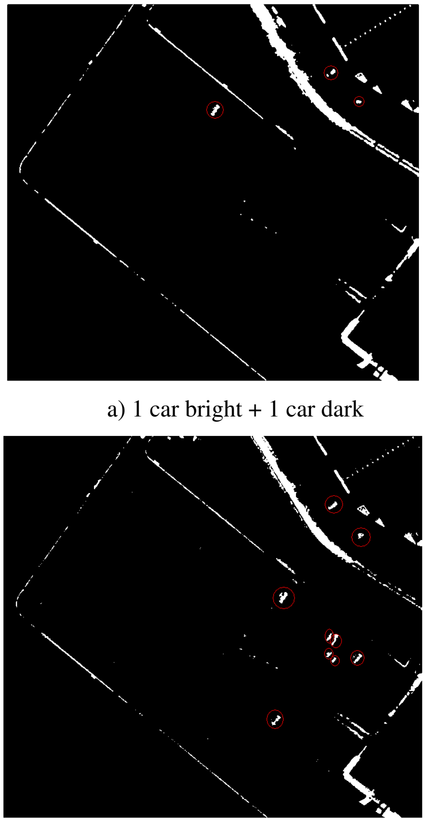
c) Expected background

Figure 12. Original satellite images and expected results.

5 . CONCLUSION AND FUTURE WORK

This paper proposes an efficient approach to reconstruct background image from satellite images, surveillance images, and street view images. It involves intelligent object detection techniques to improve the accuracy and accelerate the generation of the results. This is a generic method that can be applied and extended to a lot of GIS and ITS applications to provide pure geographic view and to preserve the privacy.

Future efforts are needed to extensively tune this method on the vehicle detection algorithms, and verify its effectiveness by using various categories of image sets taken in different scenarios. Especially, extending this approach to remove both running and parked vehicles in satellite images/street view images could be a significant contribution to the map systems such as Google Maps and Microsoft Live Search Maps.



a) 1 car bright + 1 car dark

b) Multicar bright + Multicar dark

Figure 13. Detected vehicles.

6 . REFERENCES

[1] Google.com, "Google Maps." Accessed in 2007.

[2] Microsoft.com, "Microsoft Live Search Maps." Accessed in 2007.

[3] TrafficLand.com, "TrafficLand.com." Accessed in 2007.

[4] Y. Kuno, T. Watanabe, Y. Shimosakoda, and S. Nakagawa, "Automated Detection of Human for Visual Surveillance System," in *International Conference on Pattern Recognition*, Vienna, Austria, 1996, pp. 865-869.

[5] X.-J. Tan, J. Li, and C. Liu, "A Video-based Real-time Vehicle Detection Method by Classified Background Learning," *World transactions on Engineering and Technolodgy Education,* vol. 6, pp. 189-192, 2007.

[6] D. M. Ha, J.-M. Lee, and Y.-D. Kim, "Neural-edge-based Vehicle Detection and Traffic Parameter Extraction," *Image and Vision Computing,* vol. 22, pp. 899-907, 2004.

[7] Y. Zhang, P. Shi, E. G. Jones, and Q. Zhu, "Robust Background Image Generation and Vehicle 3D Detection and Tracking," in *IEEE Intelligent Transportation Systems Conference*, Washington D.C., USA, 2004, pp. 12-16.

[8] R. O. Duda and P. E. Hart, "Use of the Hough Transformation to Detect Lines and Curves in Pictures," *Comm. ACM,* vol. 15, pp. 11-15, 1972.

[9] L.-W. Tsai and J.-W. Hsieh, "Vehicle Detection Using Normalized Color and Edge Map," *IEEE transactions on Image Processing,* vol. 16, pp. 850-864, 2007.

[10] H. Moon and R. Chellapp, "Optimal Edge-based Shape Detection," *IEEE transactions on Image Processing,* vol. 11, pp. 1209-1226, 2002.

[11] R. Alba-Flores, "Evaluation of the Use of High-Resolution Satellite Imagery in Transportation Applications," 2005.

PhD Showcase: Land Use Analysis using GIS, Radar and Thematic Mapper in Ethiopia

Haile k. Tadesse
George Mason University
Environmental Science and Public Policy
4400 University drive, Fairfax, VA
7036235751

htadess1@gmu.edu

ABSTRACT

Land degradation, and poverty issues are very common in our world, especially in developing countries in Africa. There are fewer adaptation strategies for climate change in these countries. Ethiopia is a tropical country found in the horn of Africa. The majority of the population live in rural areas and agriculture is the main economic sector. Extensive agriculture has resulted in an unexpected over-exploitation and land degradation. The project locations are Southwestern and Northwestern Ethiopia. The main objectives are to analize the accuracy of land use classification of each sensors, classification algorithms and analyze land use change. Thematic Mapper (TM) and Radar data will be used to classify and monitor land use change. Two consecutive satellite images will be used to see the land use change in the study area (1998, 2008). ERDAS Imagine will be used to resample and spatially register the Radar and TM data. The image classification for this research study is supervised signature extraction. The Maximum likelihood decision rule and C4.5 algorithm will be applied to classify the images. TM and Radar data will be fused by layer staking. The accuracy of the digital classification will be calculated using error matrix. Land change modeler will be used for analyzing and predicting land cover change. The impact of roads, urban and population density on land use change will be analayzed using GIS.

Categories and Subject Descriptors

I.4.6 [**Segmentation**]: Pixel classification; I.5.3 [**Clustering]:** Algorithm

General Terms

Algorithms, Measurement

Keywords

Land use change, GIS, Modeling, Classification Algorithm, Remote sensing

1.0 INTRODUCTION

1.1 Background

The environment is where we live; and development is what we all do to improve livelihood. As we continue to live and live better, we want to keep our environment with minimum ecological alterations. Land degradation, and poverty issues are very common in our world, especially in poor developing countries in Africa. Developing countries have fewer resources to adapt climate change and are the most vulnerable to it [20].The expansion of agriculture over the past three decades involved the cultivation of marginal areas, clearance of important natural habitats such as forests and wetlands and such conversion is a major driving force behind land degradation [19].

The majority population living in rural and urban areas is entirely dependent on the forest resources for its daily needs. In rural areas forest is a means of survival and people cut trees for firewood, construction, commercial logging and subsistence farming. There is often a conflict between local authorities and the poor community because of area closures or protected lands. Deforestation is associated with a number of economic, ecological and social aspects, which necessitate a balance between the exploitation of the direct values of the forests and the preservation of its indirect values (e.g. biodiversity, CO_2 sequestering, protection against soil erosion as well as water conservation) [18].

Nations should always assess the current state of atmosphere, land, water and biodiversity. Societies and governments should seriously acknowledge the fact that environmental degradation will diminish the potential for sustainable development. For poor people living in weak or unstable states, climate change will deepen suffering, and intensify the risks of mass migration, violent conflict, and further fragility [22]. Extreme weather conditions are having an increasingly large impact on vulnerable human communities, particularly the world's poor. Land degradation is decreasing agricultural productivity, resulting in lower incomes and reduced food security.

Africa has the potential to emerge as an exciting new center of growth in the evolving global economy [22]. To achieve this, Africa has to tackle climate change impacts by giving more attention to the issues of land degradation and poverty eradication. Investing in Agriculture and Forestry in Africa will help to combat climate change and it can also be profitable to feed the continent and the whole world. As the same time, planting trees will not have a negative effect like underground carbon sequestration or ocean Iron fertilization.

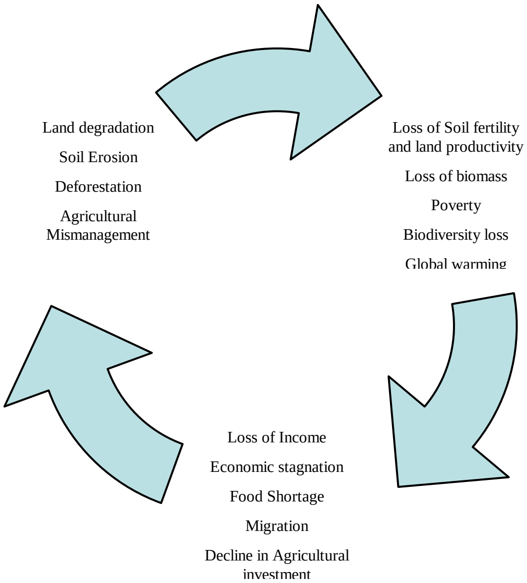


Figure 1. The interaction of Land degradation, Poverty and Global warming

For such intervention to be fruitful, it will be crucial to develop accurate quantification and verification methods of carbon sequestration in Agriculture and Forestry. Therefore, an application of remote sensing, and GIS for land-use change analysis is an essential tool to quantify the potential of African land mass on carbon sequestration and overall food security issues. Spatial information on forest degradation would enhance the effectiveness of planning development, commercial activities, and conservation activities, as well as improve local and global ecological models and carbon budget estimates [17]. Remote sensing and GIS techniques will be helpful to assess land degradation issues throughout the world. Especially in Africa, an application of Radar remote sensing can increase the accuracy of data on land use/cover change analysis.

1.2 Land Use Change and Modeling

Land use change can be caused by changes in population, market, technology, economy and income, infrastructure and environmental degradation. The drivers of land use change can be classified in to bio-physical and socio-economic factors [1]. The bio-physical factors are such as soil, topography, climate, hydrology, and others. The Socio-economic drivers of land use change are market, population, economic, political system, technology and others. There is also variability in their degree of influence from region to region and at the same time, these two categories are correlated. In general, land use change has an effect on economy, politics and environment. Both the causes and impacts of land use change can be different from one region to another.

Land use change driving factors and consequences can be analyzed by land use change models [21]. "Models can be considered as abstractions, approximations of reality which is achieved through simplification of complex real world relations to the point that they are understandable and analytically manageable" [1]. These modeling approaches can be important tools to see the trends in the future land use which may help for predicting food supply and demand in the world. Besides, environmental consequences of land use changes can be predicted to avoid catastrophic damage to society and environment.

Deforestation patterns as a result of spatially associated variables can be studied by multivariate statistical analysis [14]. Multivariate modeling helps to see the interaction between the independent variables as compared to univariate statistical analysis of deforestation [13]. In this study, early and recent deforestation can be modeled by considering the distance from towns and distance to roads respectively. A similar research study by Chomitz and Gray [2] shows habitat fragmentation and low economic returns if roads are built in areas with poor soils and low population densities. Different research studies showed that the relationship between single independent and dependent variables will ignore the importance of other factors in deforestation processes. Multiple regression analysis is one of the mostly used statistical models [1].

1.3 Landsat and Radar

Satellites have more synoptic view as compared to aerial photographs, and it is possible to see a large area of land at a given time. Satellite remote sensing provides a means to characterize the spatial patterns of vegetation changes at the landscape scale [23]. Deforestation assessment study is one of the applications in satellite industry. In a research study using remote sensing data, an overall 6.2 % land use changes from 1973-2000 occurred in the north central Appalachians [15]. The primary conversions in this study area were deforestation through harvesting and natural disturbance followed by regeneration, and conversion of forests to mining and urban lands. Such studies showed the possibility of identifying the cause of land cover change.

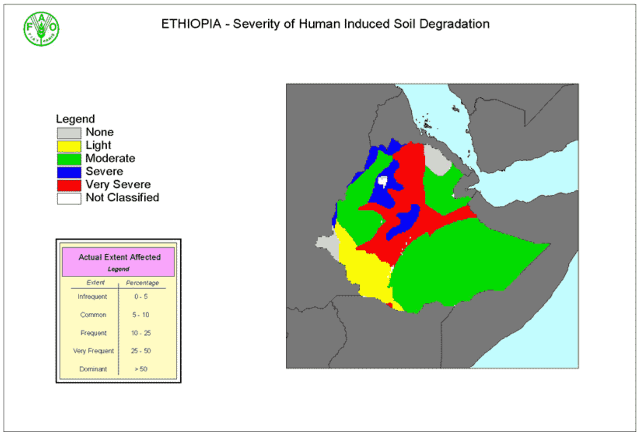
Landsat has Multispectral Scanner subsystem (MSS) and Thematic Mapper (TM) and these two scanners are sensitive to different wavelengths and this will give different spectral response for different features at a given time and space. Such variability of spectral response will help to know the ground condition of a given feature. A study using remote sensing (Landsat TM, SPOT) and fieldwork in Ecuador's eastern cordillera showed 0.58 percent annual rate of deforestation during the 1990's [11]. In this study, there is a combination of TM and SPOT and such data incorporation will help to have more accurate results. A similar study, finds that several tropical forest classes can be mapped from Landsat TM data with high accuracies [18]. However, the success of such classification is related to the existence of a comprehensive ground truth database in combination with the use of simple and intuitive methodologies.

Radar is an active sensor which send energy to illuminate the earth surface and detect the portion of back scattered portion of energy. This means they don't require another energy source like Landsat sensors. Radar is not affected by solar illumination and can be used day or night as compared to the passive sensors. Radar has the ability to penetrate atmospheric conditions like fog, hail and clouds since they have longer wavelengths. The all weather operational characteristics of radar help to collect data from different geographical areas of the world in the presence of adverse weather conditions [10]. Such advantage of radar is especially important to the case of tropics including my study area in Ethiopia.

Radar sensors receive scattered energy from the surface feature and the amount and direction of scattering is affected due to type of material, moisture content, angle of illumination and receiving, surface roughness and geometry. Microwave sensors such as radar have remained unexploited for the most part due to a lack of long-term availability of data for tropical forests, lack of appropriate bands and polarization, difficulty in interpreting data and the traditional use of optical remote sensing data by civil disciplines [16]. One of the disadvantages of space-borne system is the difficulty with the analysis of radar data [10]. This is due to the issue of single band and fixed polarization for the previous data. Now there are good polarized data available from Japanese PALSAR and RADARSAT-2. Another problem with radar data is the presence of speckle noise and this has to be corrected by de-speckling.

1.4 Ethiopia

Ethiopia has a total area of 1,127,127 km^2 and it is the third largest country in Africa. Located between 3.5^{o}N and 15.5^{o}N, Ethiopia is a tropical country found in the horn of Africa. Agriculture in Ethiopia is a source of food, raw material, foreign currency and employment. The majority of the population live in rural areas and agriculture is the main economic sector. Extensive agriculture has resulted in an unexpected over-exploitation and degradation of natural resources for many years. The topography of Ethiopia is very variable, ranging from 116 meters below sea level to 4500 meters above sea level. The majority of the population live in the highlands. The population density in the highlands and lowlands is 250 and 10 persons per km^2 respectively [3]. Land degradation in the Ethiopian highlands is mostly due to agricultural mismanagment, lack of conservation, hilly topography, intensive rainfall and low vegetation cover. The data from figure 1 shows the severity of human induced soil degradation in most parts of the highlands. In contrast, the lowlands are sparcely populated, less degraded and underutilized. This low land areas are now the center for commercial farming and land use change.



Actual Extent Affected
Legend

Extent	Percentage
Infrequent	0 - 5
Common	5 - 10
Frequent	10 - 25
Very Frequent	25 - 50
Dominant	> 50

Figure 2 Human Induced soil degradation [5].

The agro-ecology of Ethiopia ranges from desert to tropical rain forest. This diversity in agro-ecology has enabled the country to have a variety of crops and animal species. The land use of the country is divided in to arable land (12%), permanent crops (1%), meadows and pasture (41%), and forest and woodland (24%) [3]. In general, 66% of the land mass of Ethiopia is considred suitable for agriculture. Even if it is regarded as suitable land for agriculture, overexploitation with out proper land management is common throughout the country.

Agriculture is the main economic sector in the country. It is highly dependent on rainfall, which is erratic and unpredictable. Irrigation is very limited due to lack of capital, skills and knowledge. Even if the irrigation potential in Ethiopia is very high, only 1.04% (23,160 ha) is under irrigation from the total irrigation potential 2,220,000 ha in the Nile river basin [6]. Poor agricultural practices and lack of proper management approches have resulted in deforestion, soil erosion, siltation, desertification and other environmetal problems. There is frequent land redistribution and fragmentation due to population increase and land scarcity in most parts of the country. Such shortage of land resulted in local conflict, migration and deforestation. In my study area, there are also commercial farms for more than 30 years and there is on going land use change. The forest cover of the region has declined due to the combined impact of different factors and it is crucial to study land use/cover change in Ethiopia.

1.5 Statement of the Problem

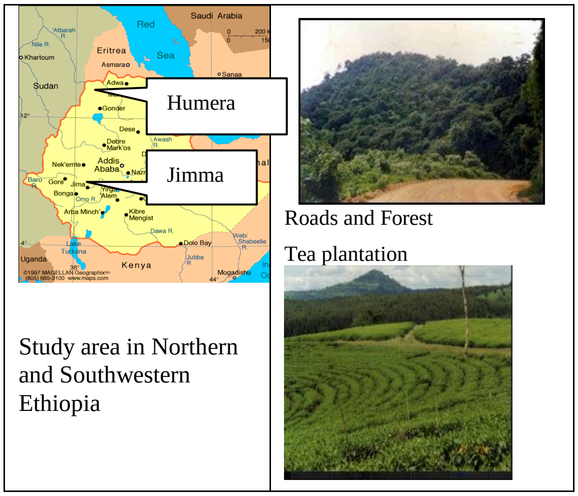
Deforestation was a common feature in Ethiopia and it is still one of its current environmental problems. The annual loss of forests in Ethiopia is estimated between 150,000 and 200,000 ha [4]. The reasons behind deforestation may vary from place to place. One study considers agricultural expansion as the main reason behind deforestation [7]. Another study in southern Wello of Ethiopia shows a 51% decline in shrub land cover due to settlement [12]. This shows that land use change drivers may vary from place to place, and it will be hard to generalize the causes of deforestation and land use change in Ethiopia. Even if issues like agricultural expansion, fire wood collection and settlements are mentioned in different studies, the impacts of road infrastructure and urbanization are not yet sufficiently addressed.

One main concern in the land use planning issues in Ethiopia is the lack of accurate data on land use/cover and deforestation. In different articles there is variability of research results and it is very hard to have a sound deforestation and/or forest cover map. If there is no accurate data on the rate of deforestation and land use issues, it will be hard to come with an appropriate plan to tackle deforestation and environmental degradation. Therefore, remote sensing and GIS techniques are important to study land use change.

Basic information concerning land use/cover is critical to both scientific and decision-making activities [8]. Such data or information is very limited in Ethiopia and in most cases decision making is very difficult. Optical remote sensing was widly used instruments to classify land-use and land-cover. In tropics and my study area, image classification using optical remote sensing results in low accuracy classified images due to cloud cover. This problem can be improved by combining the optical data with radar data. Radar holds enormous data-collecting potential for many areas, especially those often obscured by adverse weather conditions [8]. In this research Radar and Landsat TM data will be used to classify land use/cover in the study areas. Besides this the classified image will be used to analyse land use change using GIS.

2. Study Area

The project location is in Southwest (Jimma) and Northwest (Humera) Ethiopia, where the remaining major forest sites are located in Ethiopia. Besides the forest cover, the study area is home to biodiversity of coffee, wild life and other plant species. The study area is also an area with large commercial tea and oilseed plantation sites. In the study areas, there are a number of road expansions to improve the market accessibility of agricultural products and interconnect the local cities and villages.



Study area in Northern and Southwestern Ethiopia

Roads and Forest

Tea plantation

Figure 3. Study area in Ethiopia

3. Data Sources, Materials and Methods

The main data sources for this study will be satellite images. Thematic Mapper (TM) and Radar data will be used to classify the land use and monitor deforestation. Land change modeler and GIS will be used to see the current and future land use with respect to settlement/urbanization, road expansion and population density in the study area. Two consecutive satellite images will be used to see the secession of forest cover in the region (1998 and 2008). To have the same spatial resolution, the data will be re-sampled. Both images will also be geometrically rectified.

Topographic maps of the region will help to digitize the road networks and towns. Map comparison will be done based on additional secondary data and field works. ERDAS imaging will be used to classify the study area in to forest, agriculture, settlement, and pasture land. The general steps that this study follows are remote sensing image collection, field data collection, land use/land cover classification and analysis. Land use change and its correlation with the driving forces of change (roads, urban and population density) will be analyzed. Land change modeler will be used for analyzing and predicting land cover change in the study area. This model will require two consequetive land cover map for the analysis and it will also help to see the potential of transition from one cover type to another.

4. Objectives

→ To study deforestation and analyze the accuracy of land use classification in Ethiopia using Thematic Mapper and Radar sensors. Besides this, the land use classification accuracy between maximum likelihood and C4.5 algorithms will be compared.

→ The analyze and compare the classification accuracy of fusing both sensors and evaluate impact of speckle reduction and texture measures on radar data land use classification.

→ To analyze the spatial correlation of roads, urban and population density with land use change and the rate of deforestation in the region.

5. Image Acquisition, Processing, and Analysis

5.1 Field Data Collection

The primary objective of field data collection is to have training and truth sites for the supervised digital classification using Radar and TM sensors. This will be implemented with the collaboration of George Mason University (GMU), Aksum and Jimma University in Ethiopia. Before, the field trip, the principal investigator will collect all the available secondary and primary data in George Mason University. The team will use a questioner to collect land use, biodiversity, household income, source of energy, land tenure and other data from local population. The team will also collect secondary data like population, land use and soil maps from different offices. Geographic positioning system (GPS) will be used to accurately locate the overall study area and validation sites.

5.2 Data Acquisitions and Preprocessing

For this study, Radar and Thematic data from 1998 and 2008 will be used to monitor deforestation and land use change. The original data from Radar and TM will be re-sampled and spatially registered to the smallest pixel size. The images will be rectified using the GPS ground control points collected in the field or available maps. To preserve the original spectral values of the scene, the datasets will be re-sampled using nearest neighbor interpolation. To see the effects of enhancement methods on the classification accuracy of radar data, texture measures and speckle reduction will be applied.

5.3 Image Classification

The basic image classification for this research study is supervised signature extraction. This signature extraction needs prior knowledge to accurately apply and produce an acceptable accuracy. Therefore, the principal investigator will collect data and study the research area before applying the supervised signature extraction. For this classification, only four land covers mainly forest, agriculture, pasture and urban/settlements will be considered. For this research purpose, areas covered by other land cover will be considered as deforested in the final analysis. The land cover classification based on the 1998 data will be used as a base for the deforestation analysis in the area. The classification method for this research will be hard classification and it is per-pixel method.

A good digital classification needs good and representative signatures. Polygon training site selection will be applied and at least 3 training sites per land cover will be selected. The training site should be valid enough to apply to the classification algorithm and the statistical values of these will be evaluated. Such evaluation will help to avoid or reevaluate the training sites selected. ERDAS contingency test will be applied to evaluate the statistics of the signatures. After acquiring the training sites which are valid and representative for each class, the Maximum likelihood decision rule will be applied to classify the image. The contingency table produced by this rule will be applied to evaluate the accuracy of each product from the Radar and TM data. Besides this, decision tree classifier (C4.5 algorithm) will be used to classify the images and compare the classification accuracy with maximum likelihood algorithm.

Specifically for the radar data, texture measures and speckle reduction of different window sizes will be applied and compared with the original data. Window size can have a positive or negative effect depending on the intended application [9]. In a research study in Wad Medani Sudan a window size 21*21 has good producer accuracy for agriculture and bare soil as compared to the poor results for water and urban [10]. According to different studies, the window sizes affect the overall accuracy for different land cover types and different sites. In this study, the results of these different window sizes will be compared. Using these measures and the different windows, the Transformed Divergence (TD) values will be compared to the original radar data. The TD value is useful to select the best bands and band combinations for classification. For this study TD value greater than 1700 will be applied to select separability of training sites.

From time to time availability of data from different satellite sensors has increased. As a result of this it is possible to obtain data for a study area at different spatial or spectral resolution. This will help to improve the accuracy of information from land use/cover classification. Since my study area is in tropics, fusing data will help the overall accuracy of classification. In this research the TM and Radar data will be fused by layer staking. The layer staking method is the commonly used method as compared to Principal Component Analysis (PCA) and others. By layer staking different bands are combined to create a single image with multiple bands.

5.4 Accuracy Assessment

An image without accuracy assessment will be less useful since the result will be less valid for scientific purposes and other practical use by consumer. Truth/validation data collected from field will be used to analyze the accuracy of each classification from these two sensors. If these validations sites are not collected from field, other ancillary data will be used. Polygon training sites generally provide slightly lower accuracies than pixels, but texture measures will benefit [10]. For this study, three polygon validation sites for each land cover will be applied. The contingency table will help to see the producer, user and overall accuracy of the digital classification applied to the data. Such information will be used to assess the importance of each sensor to monitor deforestation status in tropical areas like the study area.

6. Final Product

The main objective of this research study is to provide an accurate data on deforestation and land use/cover change in the study area. Besides, it will provide important research findings on the type of sensors and the best classification algorithm to be used for land use change analysis. This research will identify the current cause of land use change in the area and it will recommend future policy intervention to preserve the natural resources of the region. The land cover map of the study area will be used for further research. The map will be exported to ArcGIS and this GIS layer will be used as a source of digital data base for the region.

7. REFERENCES

[1] H. Briassoulis. Analysis of Land Use Change: Theoretical and Modeling Approaches. *The Web Book of Regional Science,* Vol. 410, 2000.

[2] K.M. Chomitz, and D. A Gray. Roads, Land Use, and Deforestation: A Spatial Model Applied to Belize. *The World Bank Economic Review* 10(3):487-512, 1996.

[3] CSA. Area and production for major crop. *Statistical bulletin 227*. Central statistics agency, Ethiopia, 2000.

[4] EFAP (Ethiopia Forestry Action Programme). 1993. Draft Final Report, Volume II. The Challenge for Development. Draft Final Report. Addis Ababa, Ethiopia, 1993.

[5] FAO AGL. Land Degradation assessments in Dry Lands, Land Water division. National Soil degradation maps, 2005.

[6] FAO. Irrigation potentials in Africa: A basin approach. FAO land and water bulletin 4. Rome, Italy, 1997.

[7] Z. Gete, and H. Hurni. Implications of Land Use and Land Cover dynamics for mountain resource degradation in the northwestern Ethiopian highlands. *Mountain* Research and Development, 21: 184-191, 2001.

[8] B. N. Haack. A comparison of land use/cover mapping with varied radar incident angles and seasons. *GIScience & Remote Sensing*, 44(4):305-319, 2007.

[9] B. Haack, and M. Bechodol. Integrating multisensor data and RADAR texture measures for land cover mapping. Computers & Geosciences 26 (4):411-421, 2000.

[10] N. D Herold, B. N. Haack, and E. Solomon. Radar spatial considerations for land cover extraction. *International Journal of Remote Sensing*, 26 (7):1383-1401, 2005.

[11] B. D. Jokisch, and B. M. Lair. One last stand? Forest and change on Ecuador's eastern Cordillera. *Geographical Review*, 92: 235-256, 2002.

[12] T. Kebrom, and L Hedlund. Land cover changes between 1958 and 1986 in Kalu District, southern Wello, Ethiopia. *Mountain Research and Development*, 20 (1): 42-51, 1999.

[13] B. Mertens, and E. F. Lambin. Land-Cover-Change Trajectories in Southern Cameroon. *Annuals of Association of American Geographers*, 90(3):467-494, 2000.

[14] B. Mertens, and E.F. Lambin. Spatial modeling of deforestation in southern Cameroon. Spatial disaggregation of diverse deforestation processes. *Applied Geography*, 17(2):143-162, 1997.

[15] D. E. Napton, T. L. Sohl, R.F. Auch, and T.R Loveland. Land use and Land cover Change in the North central Appalachians ecoregion. *Pennsylvania Geographer*, 41: 46-66, 2003.

[16] S. S. Saatchi, B. Nelson, E. Podest, and J. Holt. Mapping land cover types in the Amazon Basin using 1 km JERS-1 mosaic. *International Journal of Remote Sensing*, 21**:** 201-1234, 2000.

[17] Jr. C. Souza, L. Firestone, L. M Silva, and D. Roberts. Mapping forest degradation in the Eastern Amazon from SPOT 4 through spectral mixture models. *Remote Sensing of Environment*, 87: 494-506, 2003.

[18] C. Tottrup. Deforestation in the Upper Ca River basin in Central Vietnam: A Remote sensing and GIS perspective. *Geographica Hafniensia*, C12:105-112, 2002.

[19] UNEP. Global Environment outlook (3): Past, Present and future perspectives, 2002.

[20] UNFCCC. Climate Change: Impacts, Vulnerabilities and Adaptation in Developing countries, 2007. http://unfccc.int/files/essential_background/background_publications_htmlpdf/application/txt/pub_07_impacts.pdf

[21] P.H. Verburg, P.P. Schot, M.J Dijst, and A. Verdkamp. Land use change modeling: current practice and research priorities. *GeoJournal*, 61 (4):309-324, 2004.

[22] WDR . World Development Report. Development and Climate Change, World Bank, Washington Dc, 2010. http://siteresources.worldbank.org/INTWDR2010/Resources/5287678-1226014527953/WDR10-Full-Text.pdf

[23] J. Yang, and S.D. Prince. Remote sensing of Savanna vegetation changes in Eastern Zambia 1972-1989. *International Journal of Remote Sensing*, 21**:**301-322, 2000.