A GIS METHOD FOR ANALYZING THE CRIMINAL HISTORY OF PLACE IN SÃO PAULO, BRAZIL.

Christopher Badurek
Appalachian State University
badurekca@appstate.edu

This analytical GIS method contributes to the understanding of spatial change in the context of guiding urban policy, particularly for managing crime, urban decay, and other socially-based urban problems. The change similarity data model captures the semantics required for GIS queries that address urban data. In particular, this method allows urban analysts to more easily perform a comprehensive analysis of various locations throughout the city in relation to one another. General trends of change in urban decay and demographics may be compared to trends in criminality over a multi-year period. It also simplifies comparisons made at a variety of temporal intervals to determine if there were any precipitous changes, or ‘tipping points,’ related to local demographic change or to interventions by authorities and local planning agents. The similarity model allows users to iteratively ask questions of a GIS by setting parameters through a visual query language. In this way, the user drives the data exploration process much more easily than simple mapping and GIS queries allow. A case study of change in crime rates in the central district of São Paulo demonstrates the usefulness of this method for analyzing the interrelatedness of urban place histories.

1. INTRODUCTION

This paper provides a formal description of the kinds of change similarity that may exist with regard to change in attributes among defined spatial entities. This formal description provides the theoretical basis for developing basic similarity operators for a visual query language for a GIS. In order to facilitate reasoning about process, this research discusses the semantics necessary for representing the similarity of change in the form of graphical operators and then provides description of primitives that capture the semantics of attribute change similarity. These primitives are used to represent all possible combinations of change similarity. Examples based on a case study of how these combinations are used to express queries relevant to kinds of change are then examined in detail.

In order to create a visual query language, a formal model that can be used for querying similarity of change is presented. This study is organized as follows: section two discusses the measurement of spatiotemporal change, section three presents the semantics of similarity of change relations and iconic representation, and section four presents the case study and sample queries. The last section summarizes the main contributions of this paper to spatial analysis in urban environments.

2. BACKGROUND AND THEORETICAL APPROACH

In order to enable reasoning about process, a means of measuring changes in areal units over time in relation to one another are needed. Typical SQL queries retrieve results exactly as specified. However, searching for all possible related values can be very tedious and confusing for most typical GIS users. In particular, research that addresses the problems users have with fitting their queries to relational table structures and into formal query
language expressions would enhance GIS usability [Eigenhofer and Kuhn 1999]. Similarity queries are therefore valuable since they allow a range of values to be returned as results. This approach allows answers to such questions as how to define spatiotemporal similarity, to what degree similar, and how to test a method of measuring spatiotemporal similarity of change. In performing an analysis of trends, one may use similarity operators as a means of more directly interacting with the data. The operators therefore represent the behavior the data can exhibit within this conceptual model.

**Number of Crimes 2000-2001**

![Graph showing number of crimes for four São Paulo districts]

*Figure 1. Number of crimes for four São Paulo districts.*

Information systems for socioeconomic units, such as bounded spatial entities, should support reasoning about change. Reasoning about change is a fundamental kind of question asked of data pertaining to socioeconomic units and queries concerning a unit’s change in attributes with regard to rates of change, reversibility of change, and magnitude of change should be included [Worboys 2001]. For example, consider the following case study of vehicular crime rates in the City of São Paulo (see Figure 1).

An urban planner or crime analyst may be interested in forming the following queries of crime in the central area of São Paulo: “show all districts that have the same crime attributes to a target area over a given time period” or “show all districts that have become similar to a target area over this same time period” (see Figure 2). The inverse of this operation can also be performed, allowing visualization of regions that have become completely different or dissimilar to the target area. Finally, one may also visualize all areas that have remained the same over a time period. These operators significantly shorten the cognitive process of comparing change in rates over regions and can be a potentially highly useful tool as a data filtering step in a knowledge discovery approach to exploring very large databases of crime and other socioeconomic data. It also empowers users to quickly examine trends that may vary according to linear or cyclic time frames of reference.
2.1. Modeling of Similarity Queries

In this research, similarity queries refer to parameters set to find an exact match within a database. The similarity of change data access method aims to support queries for patterns of change similar to a defined set of temporal constraints. Finding matching patterns in time series data is a typical example of the use of spatiotemporal change similarity. Examples of pattern matching in time series can be applied in a variety of domains. For example, Berndt and Clifford [1996] used ‘approximate pattern detection’ for the matching of temporal sequences, called “time series fragments,” in predator-prey relationships in ecology and trends in the stock market.

In order to further discuss the formation of a GIS that uses change similarity queries, we need to define similarity and how to measure it on various scales of measure. Similarity measures have been applied to the three main aspects of geographic data: semantic, spatial, and attributes. Semantic similarity refers to the degree of similarity among classes of geographic data, such as in classes of geographic objects. Spatial similarity refers to the geometric and topological properties of geographic objects. Attribute similarity refers to comparisons of the values associated with geographic objects. This study is unique in that it addresses the similarity of the change of objects over time.

The measurement of this kind of similarity can be addressed through the four main scales of measure [Stevens 1946]. Nominal measures generally describe classes of change, particularly distinct events that can be classified according to their attributes. Nominal classification is therefore binary and distinguishes classes based on set parameters. Examples of these kinds of change include descriptions that distinguish alterations among states of objects, such as continuous or punctuated. Nominal classifications of change are not useful for forming similarity queries except in the case of fuzzy logic or measures of the similarity of classes. Similarly, ordinal measures are simply categories placed in some sequential ordering system. Since ordinal measures focus on the ordering of values, ordinal temporal data can be represented using relative units such as before, after, or during.
Orderings can be given numeric representations based on the severity or degree of change; however these are only qualitative values and carry no mathematical operations.

Ratio measures have absolute values and are the most complete for typical kinds of mathematical operations and hence, statistical analyses. However, interval measures focus on change along time lines. Differences can be compared but not the magnitudes since there is no absolute zero. For example, time is essentially a linear scale with an arbitrary zero point which can be changed and still preserve the ordering. Interval measures describe how great the difference is between values along an ordered scale. Operations on this scale focus on meaningful differences, but not magnitudes.

With regard to spatial data, measures of similarity can be used as a data access method for large-scale spatiotemporal databases [Frank 1992], [Flewelling and Egenhofer 1999], [Goyal and Egenhofer 2001]. The means for querying similarity of change in spatial entities used here are weighted means measures. These examples work by calculating the similarity of a target to every object in a database and results are retrieved and sorted by similarity rankings.

2.2. Weighted Means Measures

The method used for the similarity of change queries is based on a measure of changes among two objects in relation to one another. A comparison of the number of attribute changes among two objects is used as a means of reasoning about overall similarity of change. Similarity used here is defined as ‘assessment of deviation from equivalence’ [Bruns and Egenhofer 1996]. A set of spatial change similarity measures is created using the weighted average method derived from entity attribute values and appended to a database. For a range of values, similarity values are derived as relative values, meaning these can be represented graphically with an x-axis composed of a range of attribute values and the y-axis a similarity range from 0 to 1. Queries can be constructed to return values that are exact matches or that fall within a threshold range. These values are scaled to 1 (same) and the degree of similarity falls along the graph in either direction.

Similarity measures can also be used for queries that use connectives such as conjunction, disjunctions, and negations. Similarity queries with conjunctions use summation measures for determining the degree of similarity among all attributes considered. In essence, this summation uses a weighted mean of the similarity of all values. For example, consider the following measure of similarity for queries based on conjunctions as defined in Nedas and Egenhofer [2003]:

$$\text{Sim}_{A_1 \ldots A_n} (O_q, O_i) = \sum_{k=1}^{n} \omega_k \text{Sim}_{A_k} (x_q, x_i)$$

Eq. 1

Consider the following query using the sample crime table: “find the similarity of all districts to District 1 for crimes in 2000” (see Table 1).

Table 1. São Paulo crime table.
This is an example of exploratory spatial data analysis in a database, including visualization through mapping to guide the process. In this case, we are essentially looking for patterns in areal data. When examining rates, exploratory approaches look at spatial variability to guide further analysis. However, the usefulness of raw rates are limited due to the fact that counts are relative to the over population per spatial unit. Consider the results from this query using the following table:

Table 2: District Similarity Rankings.

<table>
<thead>
<tr>
<th>District</th>
<th>Name</th>
<th>Similarity Value</th>
<th>Total 2000</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Sé</td>
<td>1.00</td>
<td>141</td>
</tr>
<tr>
<td>4</td>
<td>Consolação</td>
<td>0.94</td>
<td>150</td>
</tr>
<tr>
<td>5</td>
<td>Aclimação</td>
<td>0.79</td>
<td>171</td>
</tr>
<tr>
<td>78</td>
<td>Jardins</td>
<td>0.77</td>
<td>174</td>
</tr>
<tr>
<td>77</td>
<td>Santa Cecília</td>
<td>0.76</td>
<td>107</td>
</tr>
<tr>
<td>2</td>
<td>Bom Retiro</td>
<td>0.72</td>
<td>102</td>
</tr>
<tr>
<td>3</td>
<td>Campos Eliseos</td>
<td>0.54</td>
<td>76</td>
</tr>
<tr>
<td>8</td>
<td>Brás</td>
<td>0.39</td>
<td>227</td>
</tr>
<tr>
<td>6</td>
<td>Cambuci</td>
<td>0.00</td>
<td>489</td>
</tr>
<tr>
<td>12</td>
<td>Parí</td>
<td>0.00</td>
<td>341</td>
</tr>
</tbody>
</table>

This query examines the degree of similarity of all districts to District 1 with respect to values scaled to one. Note that although similarity values may be close in value they can be distant from each other in raw value. For example, District 5 and District 77 have very close similarity values, but have numbers above and below that of District 1, the referent value. The results of the query are returned in the following table, ordered by rank similarity (see Table 2). In this case, District 4 is most similar with a similarity value of 0.94.

3. SEMANTICS PRESENT IN SIMILARITY OF CHANGE RELATIONS

3.1. Visual Operators

Research on visual query languages can be described as a branch of research on visual query systems [Catarci et al. 1997]. These kinds of information systems are oriented towards novice users or those with limited facility with forming complex queries with traditional query languages such as SQL. A major advantage of these using visual representations is that they can be used to as a basis for visual query languages that can shorten the process of building queries that may be too complex for the average user to construct. These query languages can be described as a tool that facilitates user interactions with data that may otherwise be too complex or difficult to make comparisons with [Tauber et al. 1994]. Visual query languages are often built from primitives, which describe the basic kinds of queries that represent some relationships or objects in the domain of interest. For example, primitives can be combined to form much more complex and powerful queries that more
efficiently and effectively express the semantics of complex queries [Hornsby and Egenhofer 1997], [Bonhomme et al. 1999]. In order to build a visual query language that represents similarity queries, an iconic representation for similarity relations must be formalized.

In this paper, a set of basic operators are systematically described as a foundation for the methods that may be performed on spatial objects in regards to temporal comparisons. These basic operators are then extended to a complete set of change operations. The purpose of describing these operators is to provide a foundation for a basic spatiotemporal query language that captures the semantics of reasoning about temporal similarities. The case study focuses on areal objects, where similarity values are the attribute values or states that are of primary interest.

3.2. Similarity Primitives

In this research, primitives refer to the basic operators that form the foundation of a visual query language for change. Here, primitives are binary and direction is a description of the values or attributes associated with a behavior. These primitives refer to values of the states of each of the objects in the domain. The primitives are therefore described as one of the following:

Let \( O = \{ O_1, \ldots, O_n \} \)

\[ \exists t \in O, s(T \geq R) \]

\[ \exists t \in O, s(T \leq R) \]

This means that similarity values can hold one of the following three values:

1. **same**, \( R = T \), meaning the two objects have the same similarity value;
2. **greater than**, \( R > T \), meaning the two objects are similar and the target object has a greater value in terms of the attributes being compared;
3. **less than**, \( R < T \), meaning the two objects are similar and the target object has a lesser value in terms of the attributes being compared.

In this sense, the relationship among two states can have similarity values of 0, 1, or a value in between. A value of 0 represents complete difference, or absolute dissimilarity. A value of 1 represents absolute similarity, or the two objects are of identical similarity. The similarity value therefore represents that the target has a difference metric that is lesser than the referent, but its attribute value is of a greater value.

3.3. Visual Operators

Visual languages are constructed to take advantage of spatial concepts inherent to their iconic representations which give cues to the user of their meaning. For example, these primitives are represented using the following graphical conventions: black represents the referent object, white the target object, and the spatial relationship among the two objects, the similarity values in regard to each other over time 1 to time 2. Consider the following example using sample data values (see Figure 3).
The grey object refers to values that are the same or highly similar. This is a reference to Object 1 at $t_1$, so Object 1 at $t_2$ is compared to Object 1 at $t_1$. The change values of Object 1 (0.7 to 0.7) are compared to the changes values of Object 1, (0.7, 0.3), meaning Object 1 changes less than Object 1. These therefore refer to change over the time intervals, $t_1$ to $t_2$ and $t_2$ to $t_3$, respectively. The next section describes how these primitives are used to develop operators for querying databases.

Figure 4. Similarity Operators Iconic Representation.

Figure 4 describes all nine relations in the form of a visual language. The first five (a-e) are primitive operators as the remaining four are only the inverse of the first four in terms of possible relations among two objects. Therefore, there are five primitive similarity operators present in Figure 4: Face (a,i), Converge (b,h), Invert (c,g), Diverge (d,f), and Hold (e).

4. CASE STUDY

To present the utility of the change similarity approach, consider the following map showing the similarity of counts of crimes per unit over time in the centrally located Sector One of São Paulo (see Figure 5). While crime is widespread throughout the city, it is mainly concentrated in specific areas. These areas tend to be associated in
places with high rates of poverty or in central areas which attract criminal activity. For example, 2/3 of the 96 districts are below what are generally considered acceptable living conditions from a developing country perspective (Câmara et al. 2004). These regions are generally on the outer edge of the inner core but many property crimes occur in the center of the city.

![Map of similarity of change in vehicular robbery](image)

**Figure 5.** Change similarity relations in Sector 1.

Examination of these patterns leads to an intuitive sense of similarity of change among spatial units. In this case, District 1 is highly similar to District 3 and generally dissimilar to Districts 6 and 12. However, there are points along the distribution of values that show similar trends in either an increasing or decreasing direction. Yet, examining the table of values yields more description of the change that is occurring (see Table 3). Consider the relation between District 1 and Districts 3, 4, and 12 respectively.

**Table 3. Similarity of Change Values:**

<table>
<thead>
<tr>
<th>District</th>
<th>Name</th>
<th>Final</th>
<th>Direction</th>
<th>Count r1</th>
<th>Count r2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Sé</td>
<td>1.00</td>
<td>neg</td>
<td>141</td>
<td>127</td>
</tr>
<tr>
<td>5</td>
<td>Abimação</td>
<td>0.92</td>
<td>neg</td>
<td>171</td>
<td>148</td>
</tr>
<tr>
<td>77</td>
<td>Santa Cecília</td>
<td>0.91</td>
<td>neg</td>
<td>107</td>
<td>97</td>
</tr>
<tr>
<td>4</td>
<td>Consolação</td>
<td>0.90</td>
<td>neg</td>
<td>150</td>
<td>147</td>
</tr>
<tr>
<td>78</td>
<td>Jardins</td>
<td>0.89</td>
<td>neg</td>
<td>174</td>
<td>161</td>
</tr>
<tr>
<td>2</td>
<td>Bom Retiro</td>
<td>0.80</td>
<td>pos</td>
<td>102</td>
<td>119</td>
</tr>
<tr>
<td>8</td>
<td>Brás</td>
<td>0.67</td>
<td>neg</td>
<td>227</td>
<td>225</td>
</tr>
<tr>
<td>3</td>
<td>Campos Elíseos</td>
<td>0.64</td>
<td>pos</td>
<td>76</td>
<td>101</td>
</tr>
<tr>
<td>6</td>
<td>Cambuci</td>
<td>0.53</td>
<td>neg</td>
<td>489</td>
<td>367</td>
</tr>
<tr>
<td>12</td>
<td>Pari</td>
<td>0.53</td>
<td>neg</td>
<td>341</td>
<td>249</td>
</tr>
</tbody>
</table>

In this case, the similarity of change among these three is represented in the similarity query language in the following manner:

*Similarity of District 1 and District 3*
The transition of Districts 1 (D1) and 3 (D3) is described by the Converge operation. The referent (D1) is decreasing while the target (D3) is increasing until they have a similar rate at time 2. The change in values of these objects is decreasing for D1 (141 to 127) and increasing for D3 (76 to 101).

Similarity of District 1 and District 4

The transition of Districts 1 (D1) and 4 (D4) is described by the Diverge operation. The referent (D1) is decreasing while the target (D4) is also decreasing at time 2. The change in values of these objects is decreasing for D1 (141 to 127) and slightly decreasing for the target D4 (150 to 147). One is either absolute or set to a value range of 10, where within a range of 10 of the target values are therefore represented as equivalent.

Similarity of District 1 and District 12

The transition of Districts 1 (D1) and 12 (D12) is described by the Pace operation. The referent (D1) is decreasing while the target (D12) is decreasing similarly at time 2. The change in values of these objects is decreasing for D1 (141 to 127) as well as for D12 (341 to 249).

These visual operators describe the changes in counts in relation to the referent object. Overall, similarity values for these districts indicate that D4 (0.90) is changing most similarly to D1, followed by D3 (0.64) and D12 (0.53). This is evident by examining the table as D1 and D4 are most similar in rate of change while the others are substantially different. For example, D12 has a very large crime value and D3 is changing similarly, but in the opposite direction (positive) of D1. Queries would be written using these visual operators for precise retrieval of these similarity values.

A method that enhances reasoning about change is to integrate similarity queries with visualization. For example, a relatively simple but effective approach is to couple matrices with queries. With the similarity of change approach, one can represent the similarity of change over time for easier access to identifying overlapping temporal units. Consider the following matrices that represent the similarity of District 1 to all other districts over monthly and quarterly temporal intervals (Figures 6 and 7). In the example of quarter intervals, clearly time units 1 and 7 have the most districts similar to District 1. Districts 4 and District 2 also appear to be changing most similarly to District 1.

![Similarity matrix: city districts (left) and quarterly temporal units (top).](image-url)
5. CONCLUSION

The results of this study are a conceptual model of similarity of change and a theoretical framework for a visual query language. This paper has presented a set of query operators that may be used for querying the similarity of change with a visual query language. First, primitives that form the basis for a visual query language are presented. The case study has presented the similarity of change query language and its use to visualize kinds of change described by each element of this visual language.

This research has sought to make it easier for GIS users to reason about change by using qualitative measures rather than metric queries. As well-discussed in the geographic information science literature, SQL is limited as a means of adequately expressing qualitative and complex spatiotemporal queries. By enhancing the expressiveness of SQL-based query languages, the capabilities of GIS may be enhanced such that they may support exploratory data analysis of urban environments. Therefore, recommendations based on these findings could be used to inform future development of querying capabilities to support urban and regional analysis.

REFERENCES


