Chapter 6

RESULTS

This chapter presents three use-cases of our methodology for modeling and mining spatial data using the proposed graph-based representations. For this purpose, we used two spatial databases as our test domains. The first database contains data related to a population census from the year of 1777 in Puebla downtown and the second one is a database storing data from the Popocatépetl volcano.

The use-cases described were implemented based on the following premises: evaluating the graph-based proposal for modeling and mining spatial data, evaluating the limited overlap feature, and evaluating the discovered knowledge with the support of a domain expert. Therefore, the three use-cases presented in this chapter were implemented based on the following methodology:

1. Selection of the spatial layers to work with.
2. Selection of the spatial relations that will be validated among the spatial objects.
3. Selection of the non-spatial attributes that will be related to the spatial objects in the graph.

4. Mining the graph using the no overlap, standard overlap and limited overlap features.
5. Evaluation of discovered patterns.

6.1 Population census from the year of 1777 in Puebla downtown

As we have already mentioned, our first test domain is a spatial database containing data of a population census from the year of 1777 (see Chapter 5 for more details). Figure 6.1 shows a fragment of the "*chpuebla*", "*chiglesia*" and "*chrio*" spatial layers used in the use-cases.

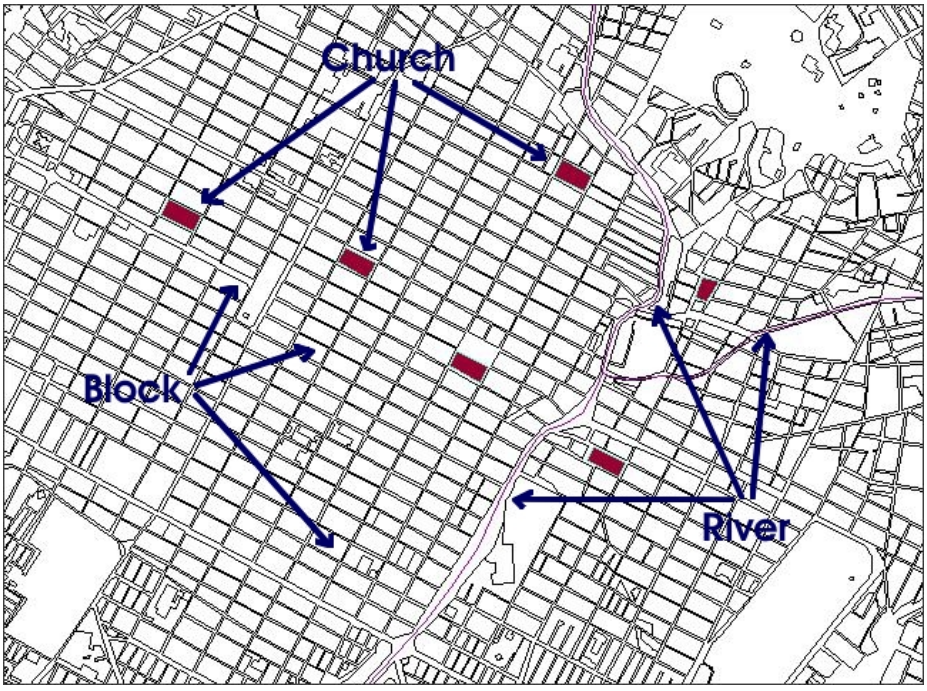


Figure 6.1. Population census from the year of 1777 in Puebla downtown.

The *chpuebla* spatial layer (shown in white color) represents blocks in Puebla downtown; this layer is related to a population census from the year of 1777 as non-spatial data. The *chiglesia* layer (shown in red color) contains representative churches for each parish in the zone. The *chrio* layer (shown in green color) represents a river crossing Puebla downtown.

It is important to remark that a parish is a spatial object grouping several blocks in the zone. Each parish has a church as its agglomerative element (people used to live around a church). Figure 6.2 shows the six parishes (each shown in a different color) and their representative church:

1. El Sagrario.
2. San José.
3. San Marcos.
4. San Sebastián.
5. Santa Cruz.
6. Santo Angel.

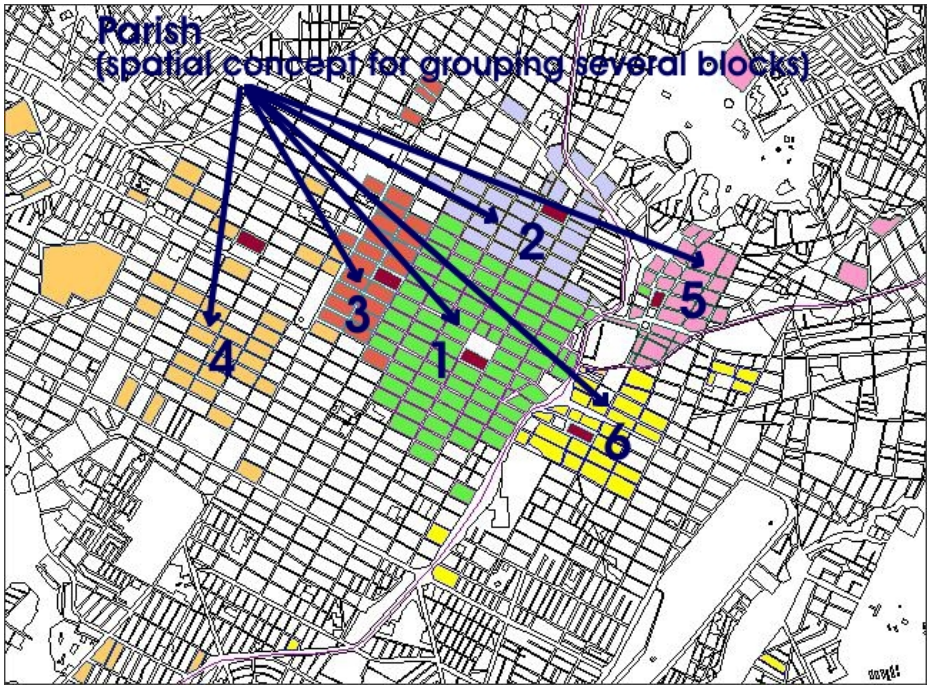


Figure 6.2. Parishes in Puebla downtown in the 1777 year.

All use-cases presented in this section were developed using all graph-based models (currently five models). However, the description of the generated results in this first test domain corresponds only to model #2 (the model is named single replication of relation types, complete information). Some of the discovered patterns were used by the domain expert to validate facts already known (i.e., distribution of the population in the census) and other allows him to know unknown relationships among spatial objects and non-spatial data (attributes) in the census (i.e., common characteristics of people living along the two borders of the river crossing Puebla downtown). In Section 6.2 we present a comparison among the generated results by each proposed model using a Popocatépetl volcano database.

6.1.1 Use-case: El Sagrario

Suppose we want to know what people have in common in the spaces within a radius of 150 meters from the representative church in parish #1 (El Sagrario). Our experiment will be focused to find regularities related to the following issues:

- Number of habitation spaces in a house.
- Members of a family.
- Type of family.
- Ethnic group of each family member.

The guideline to select a radius of 150 meters from the church is that this value allows us to include in our sample dataset at least one block in all directions around the church as we show in Figure 6.3. Thus, by using the prototype system we selected the *chpuebla* and *chiglesia* spatial layers. Once we selected the spatial layers to work with, the following steps are to select the spatial relation(s) to be validated among the spatial objects and the non-spatial attributes that will be related to the spatial objects in the graph. The selected parameters were the following:

- Spatial layers: *chpuebla* and *chiglesia*.
- Pivot: representative church in parish #1.
- Spatial relation: distance.
    - Value: 150 meters within a radius from the representative church to the blocks.
- Spatial graph-based model: model #2.

Figure 6.3. Blocks 150m. from representative church, parish "El Sagrario".

The generated graph was composed of 24,167 vertices, and 24,166 edges according to the proposed graph-based model. Figure 6.4 shows a fragment of the graph.

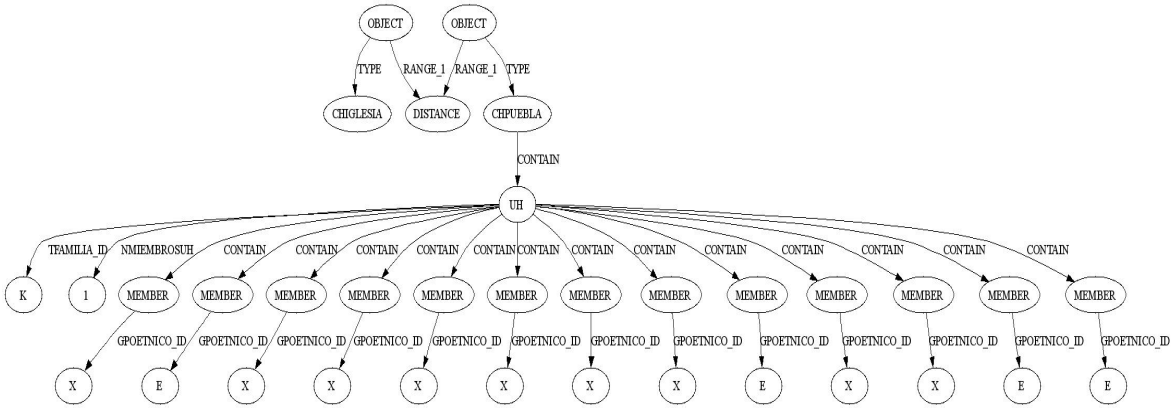


Figure 6.4 Example of a generated graph in the use-case "El Sagrario".

This graph was used as the input to the Subdue system. For the experiment we used the following Subdue's parameters:

- Predefined substructure: yes (we used "*UH CONTAIN MEMBER*" since these elements are grouping components in our graph-based representation for the non-spatial data of the census). See Section 5.1 for more details.
- Overlap: yes.
- Limited overlap: no/yes (first, we used standard overlap; next, we used limited overlap).

Once Subdue completed the mining process, the generated results (patterns) were evaluated by the domain expert. Figure 6.5 shows three examples of discovered patterns using the standard overlap option. The expert's opinions were based on the following issues: the patterns are based on the population distribution schema in the census from the year of 1777 (large population inhabits in parish #1). 65% of the population, in this area, did not given its ethnic group, this can be interpreted from a demography history perspective, as a possible dissolution of the racial element for grouping people (creation of groups or classes) in benefit of alternative grouping parameters such as salary, family networks (how they lived and whom they lived with), consumption levels, type of house. 16% said to be "Spanish". People lived based on the model "*Jefe con Familiares y Agregados*".

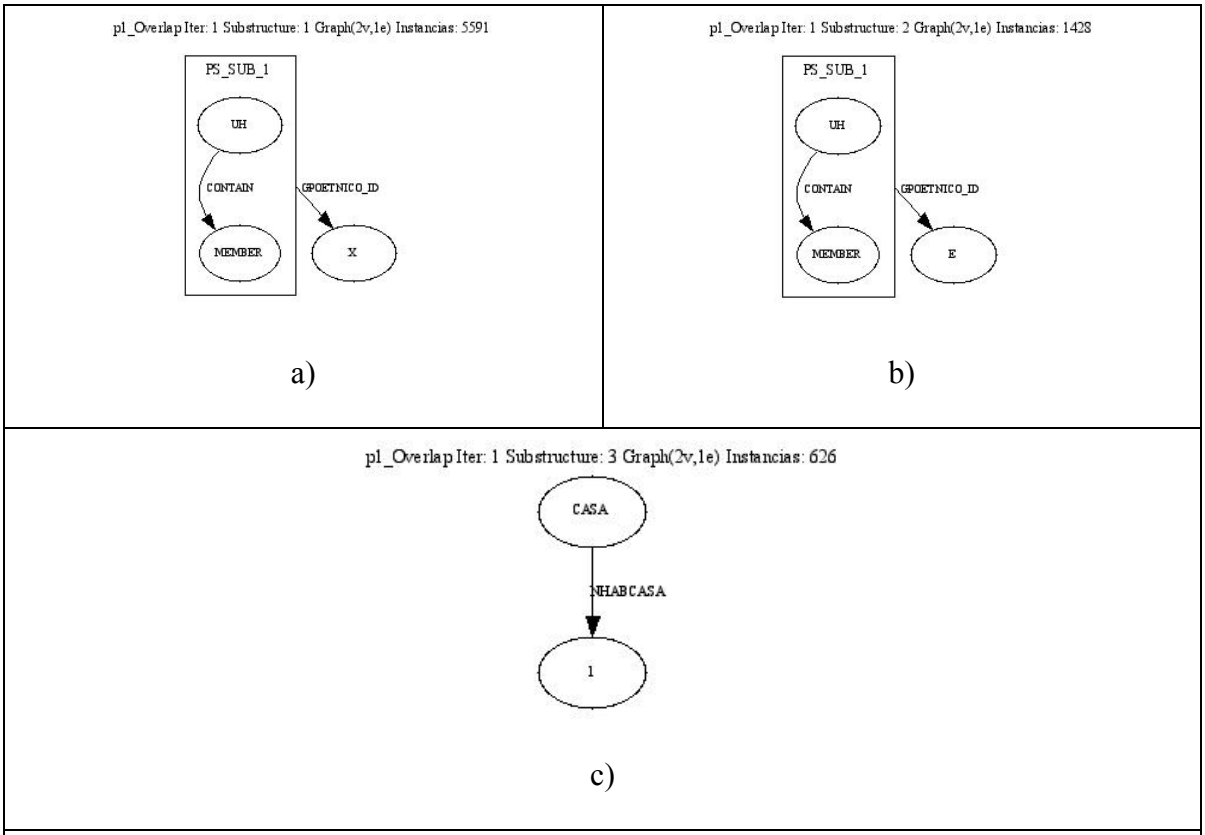


Figure 6.5. Examples of discovered patterns by using standard overlap in use-case "El Sagrario" (1).

We can see that the patterns shown in the figure are composed of spatial data (i.e., a vertex representing a house) and non-spatial data (i.e., a vertex and an edge telling us the major ethnic groups in the census). Although those patterns do not contain any reference to spatial relations we must to note that a spatial relation was used to define the dataset for the mining phase (i.e., 150 meters within a radius from the representative church to the blocks). They do not appear in the discovered patterns but they are part of the input graph. An explication to this situation is that most common characteristics (repetitions in the input graph), in this example, occur in the non-spatial data describing the spatial objects (i.e., the ethnic group "*X*- Undefined" and "*E*- Spanish"). However, these facts represent the strength of our proposal since our basic idea is to represent spatial data, non-spatial data and spatial

relations among the spatial objects as a unique dataset, mine it together, so we can discover patterns involving these elements at the same time.

The next step in our experiment was to evaluate our limited overlap proposal in Subdue. We stated three motivations to propose the new algorithm: a search space reduction, time processing reduction and specialized overlapping pattern oriented search. Therefore, in order to demonstrate that by using the new approach we obtain those benefits we implemented the following example. For the limited overlap test, we allow overlap only in vertices representing the ethnic group of the family members. These vertices are used to guide the overlapping pattern oriented search.

In this example, the discovered patterns by using the standard and limited overlap features were slightly different. Figure 6.6 presents three examples of discovered patterns using the limited overlap. For example, both cases reported as the first substructure a pattern telling us "Undefined" is the predominant ethnic group in the dataset. The second discovered substructure, for both cases, tells us that "Spanish" is the next predominant ethnic group. In the case of the third substructure, using standard overlap Subdue reported a pattern related to the number of habitation spaces in a house, but through limited overlap Subdue found a relationship among the "Undefined" ethnic group and the family type "*Jefe con Familiares y Agregados*". An interpretation of these results is that limited overlap reports in all discovered substructures a vertex representing ethnic group because we oriented the search over this type of vertices when we specified that vertices representing ethnic group were allowed for overlapping.

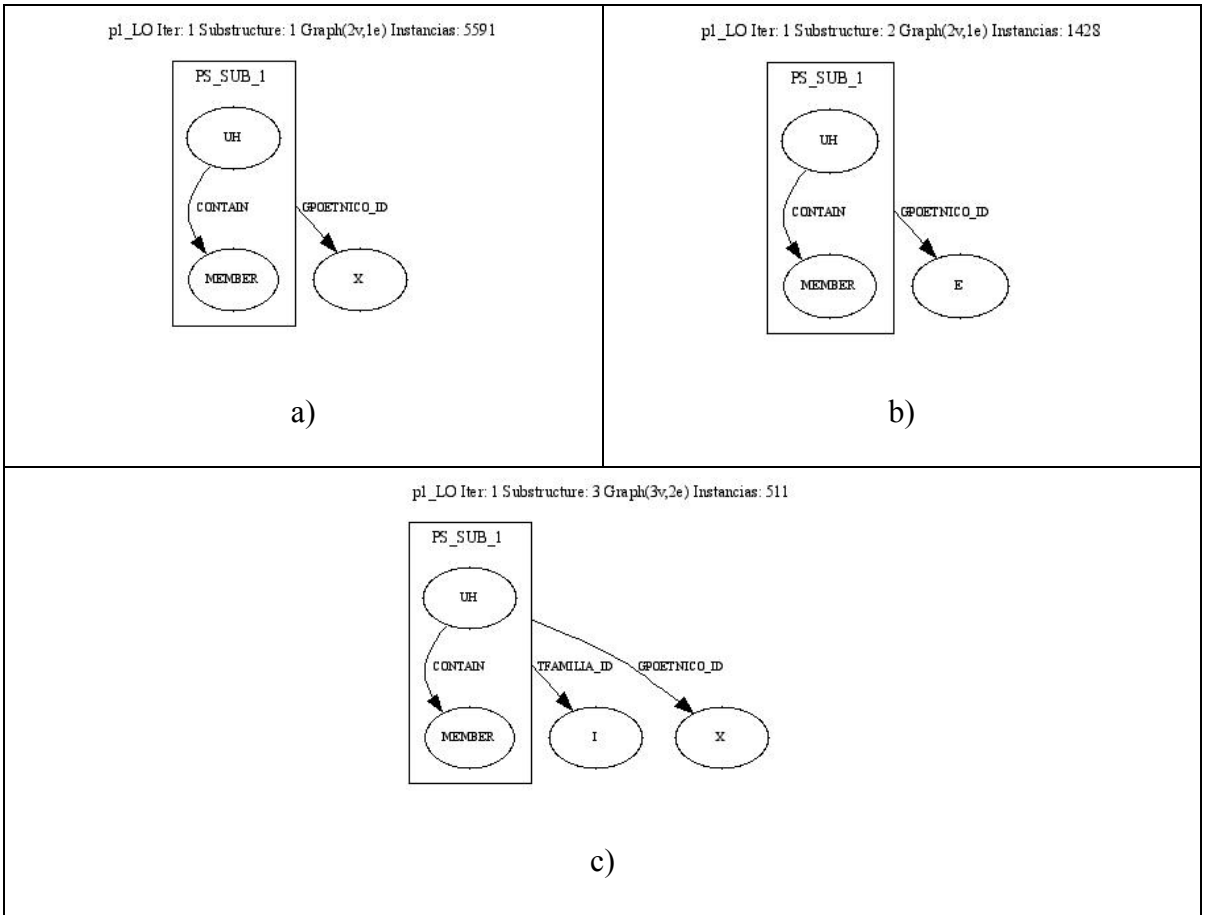


Figure 6.6. Examples of discovered patterns by using limited overlap in use-case "El Sagrario" (1).

The next step in our experiment was to evaluate the objective of a processing time reduction by using the limited overlap feature over the standard overlap. Figure 6.7 presents the processing time comparison chart for this experiment. In the figure, the time taken by using standard overlap is shown in pink color (193,426 seconds). On the other hand, the processing time taken by using the limited overlap option is shown in yellow color (36,042 seconds). As we can see, we obtained a time reduction gain of 81.37% using our proposed limited overlap approach.

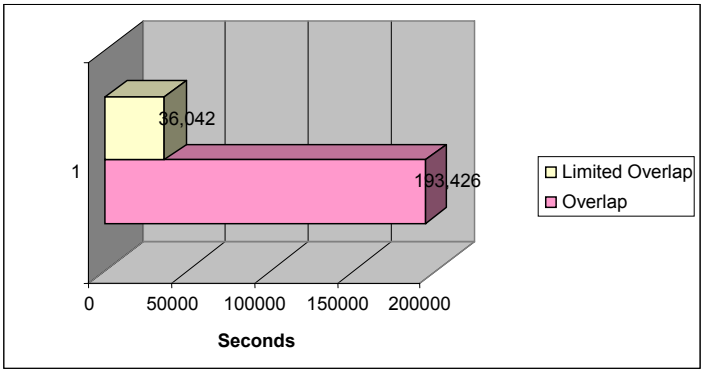


Figure 6.7. Processing time standard vs. limited overlap: use-case "El Sagrario" (1).

Now, we are going to modify slightly our study zone to show the way we can work with two or more spatial relations at the same time. By using the prototype system the user can select one or more spatial relations that will be validated among the spatial objects. For instance, it is possible to select two spatial relations belonging to the same spatial relation type (i.e., topological) or to different spatial relations types (i.e., one topological and one distance relation). Suppose we want to search for regularities about people and habitation spaces in a radius of 150 meters from the same representative church in parish #1, but this time, we only want to evaluate blocks located on the North side as shown in Figure 6.8. Our experiment will be focused to know common regularities over the same issues as in the previous test.

By using the prototype system we selected the following parameters:

- Spatial layers: *chpuebla* and *chiglesia*.
- Pivot: representative church in parish #1.
- Spatial relation: distance.

- Value: 150 meters.

- Spatial relation: direction
  - Value: North.

- Spatial graph-based model: model #2.

The generated graph was composed of 12,021 vertices, and 12,026 edges according to the proposed graph-based model.

Figure 6.8. Blocks 150m. North side from representative church, parish "El Sagrario".

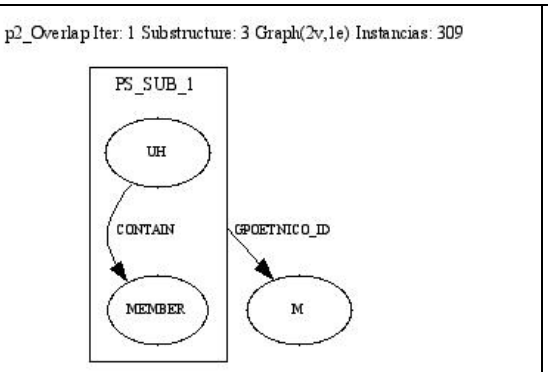
As in the previous example, the graph was used as input to Subdue. For the experiment we selected the following Subdue's parameters:

- Predefined substructure: yes (we used "*UH CONTAIN MEMBER*" since these elements are grouping components in our graph-based representation for the non-spatial data of the census). See Section 5.1 for more details.
- Overlap: yes.
- Limited overlap: no/yes (first, we used standard overlap; next, we used limited overlap).

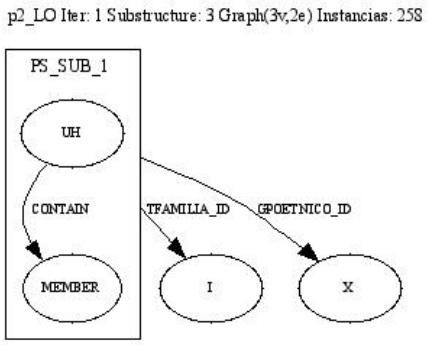
Once Subdue completed the mining phase, the generated results were evaluated by the domain expert. The expert found the following facts: the predominant ethnic group in the area remains as "Undefined" because of the proximity to the parish center and the continuous racial and social interchange from the North side. As consequence, they share the same family type structure. The "Spanish" population is the most important (15%) and the next one is "Mestizos" (7.13%). The family nucleus which includes "other people living with" (added people) employs "Mestizos" as subordinated workers (i.e., waiters and salesmen). It is important to remark that they do not employ "Indígenas" (maybe because they do not speak Spanish and their limited cultural level).

This is the domain expert evaluation but we also needed to evaluate how the new limited overlap feature worked. Therefore, the same experiment was performed using the standard and then the limited overlap. For limited overlap, the vertex allowed for overlap was the same as in the previous test. The discovered patterns by using standard and limited overlap were very similar to those obtained in the first test. In fact, the first and second reported substructures were the same although the third one was different (see Figure 6.9). By using

standard overlap Subdue reported "Mestizos" as the third predominant ethnic group. Limited overlap reported a relationship among the "Undefined" ethnic group and the family type "Jefe con Familiares y Agregados". This last pattern was the same one reported in the previous test.



a)



b)

Figure 6.9. Examples of discovered patterns in the use-case "El Sagrario" (2)

Figure 6.10 shows the processing time comparison chart by using the standard and limited overlap features. In this figure, the time taken when using the standard overlap feature is shown in pink color (30,532 seconds) while the processing time taken when using limited overlap is shown in yellow color (2,471). It is important to note that by using our new overlap approach we obtained a time reduction gain of 92% in our experiment.

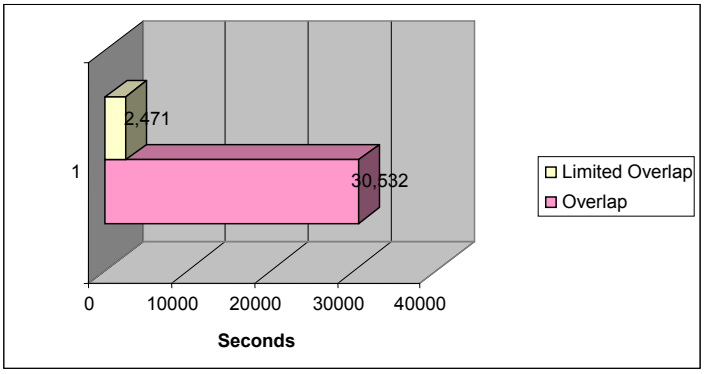


Figure 6.10. Processing time standard vs. limited overlap: use-case "El Sagrario" (2).

6.1.2 Use-case: People living along the borders of the river crossing Puebla downtown

As we mentioned at the beginning of the chapter, a river crosses Puebla downtown. Suppose that we are interested in knowing characteristics about family types and ethnic groups of people living along the borders of the river.

Therefore, we defined as our study area all blocks located at most 50 meters from the borders of the river as shown in Figure 6.11. We selected this distance since it allows us to select at least one block along the entire border of the river. We used the following parameters in our use-case:

- Spatial layer: *chpuebla* and *chrio*.
- Pivot: the river.
- Spatial relation: distance.
    - Value: 50 meters.

- Spatial graph-based model: model #2.

The generated graph was composed of 16,597 vertices, and 16,596 edges according to the proposed graph-based model.

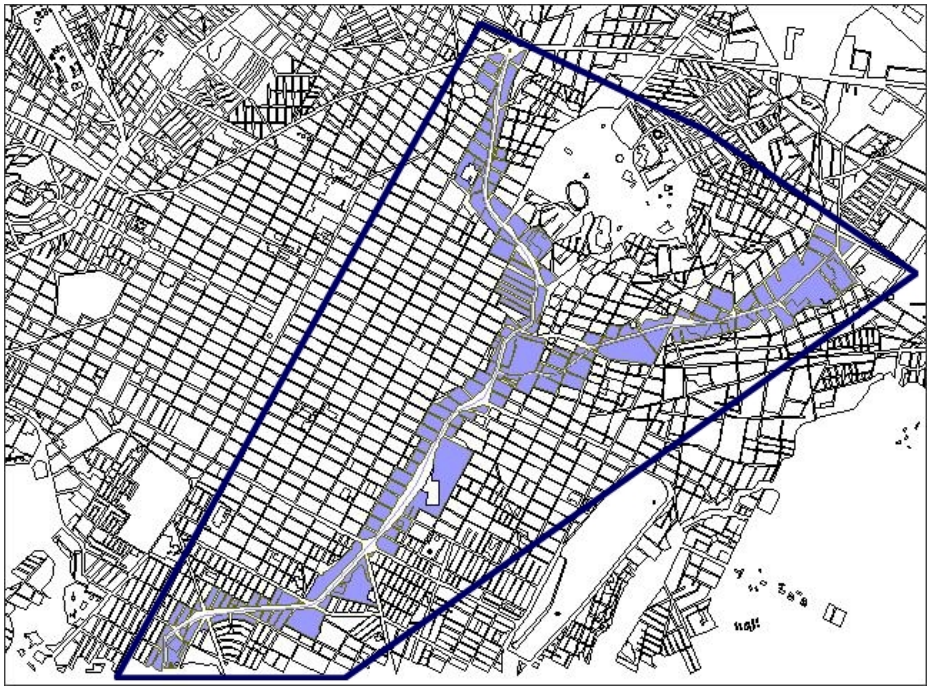


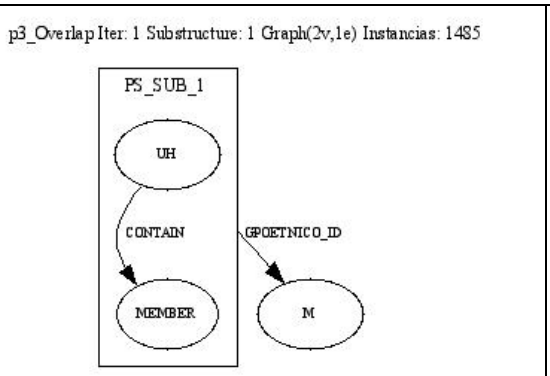
Figure 6.11. Blocks 50m. from river crossing Puebla downtown.

The created graph was used to feed the Subdue system. For this test we selected the following Subdue's parameters:

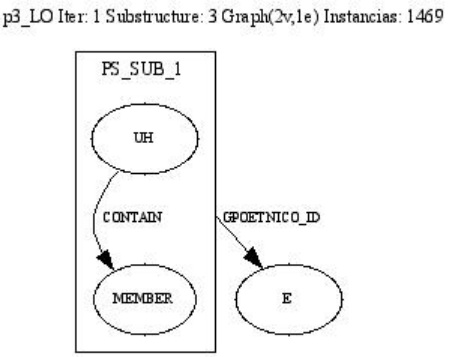
- Predefined substructure: yes (we used "*UH CONTAIN MEMBER*" since these elements are grouping components in our graph-based representation for the non-spatial data of the census). See Section 5.1 for more details.
- Overlap: yes.

- Limited overlap: no/yes (first, we used standard overlap; next, we used limited overlap).

Figure 6.12 shows two examples of discovered substructures in this use-case. Our domain expert evaluated the generated results focusing in the following issues: there is a modification for the population agglomerative criterion on the East side ("San Francisco", "El Alto Xonaca", and "Los Remedios" neighborhoods) and on the West side ("San José", "El Sagrario" and "El Carmen" neighborhoods) of the river. The "Mestizos" is the predominant ethnic group (24.5%); the "Spanish" is the next one (almost has the same percentage). This pattern may outline that the "Mestizos" ethnic group played the role of intermediator between the "Spanish" and "Undefined" groups on the West side, and between the "Spanish" and "Indígenas" on the East side.



a)



b)

Figure 6.12. Examples of discovered patterns in use-case "people around the river crossing Puebla downtown"

The previous results were generated using standard overlap, but we needed to compare them with the results obtained using limited overlap, so we used the same parameters as in

the previous test. We proposed to allow overlap only in vertices representing the ethnic group.

It is important to mention that, if our overlap label list has many elements, it could be more efficient (concerning to processing time) to use standard overlap because of the overhead that results from the validation process. Remember that when we use the limited overlap feature, each time that an overlap among instances of a substructure is detected, the overlap is evaluated in order to know if it is allowed or not.

In this experiment we noted that the patterns discovered using the standard and limited overlap approaches were the same but the processing time taken by each of them for the mining task was slightly different. Figure 6.13 presents the time comparison chart for the experiment showing the time taken when using standard overlap in pink color (15,572 seconds) and the time taken when using limited overlap in yellow color (14,021 seconds). Limited overlap required a lower time to find the same patterns than standard overlap.

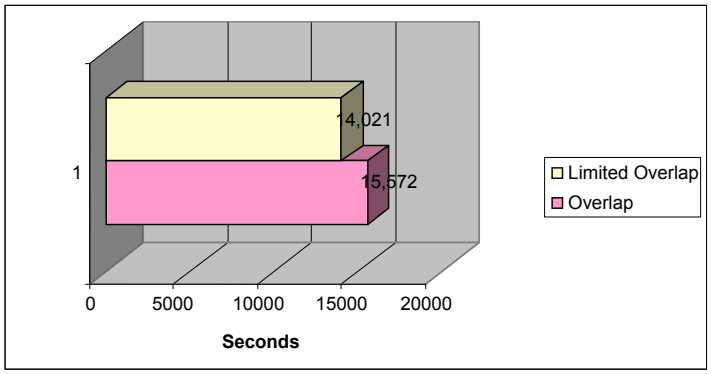


Figure 6.13. Processing time standard vs. limited overlap: use-case "people around the river crossing Puebla downtown".

The two use-cases presented in this section show the functionality of our proposal for modeling and mining spatial data using a graph-based representation. The discovered patterns in the population census from the year of 1777 were analyzed by a domain expert. Some of these patterns allow the user to validate facts already known. For instance, the predominant ethnic groups classified by parish, and population distribution according to the social status in the zone. But other patterns allow him to know implications, previously unknown, among spatial concepts and non-spatial attributes in the census. For instance, racial and economic interchange among people in the parishes, which are the common characteristics of population living along the borders of the river crossing downtown (on the West and East sides), social structure according to the ethnic group, common regularities among the family type and habitation space.

Processing time comparison among standard and limited overlap features was other topic evaluated in these use-cases. We presented time processing comparison charts showing the time reduction gain obtained by using the new approach. We also demonstrated that by using limited overlap we can orient the search over substructures (patterns) containing elements that in our domain may represent relevant issues. For instance, in the year of 1777 the ethnic group represented a significant element to know characteristics about a family and their habitation space.

Next section presents a use-case using a Popocatépetl volcano database. The objective in this illustrative use-case is to evaluate/compare the generated results by each proposed graph-based model.

6.2 Popocatépetl volcano

In Section 4.3 we presented a preliminary use-case showing the applicability of our methodology using a Popocatépetl volcano database. This section presents an extended use-case employing the same database. First, we describe the study area, next the method used to define the dataset and the parameters for the spatial data mining processes, and finally the generated results.

As already mentioned, the database contains data related to several issues in the Popocatépetl zone such as settlements, rivers, and evacuation roads. Figure 6.14 shows a fragment of the Popocatépetl volcano database. For the experiments we will use three spatial data layers. The first one is the roads layer (representing the roads in the area), the second one is the rivers layer (representing the rivers in the area), and finally the third one is the settlements layer (representing population areas). To illustrate the use-case we have delimited a study zone as shown in the figure.

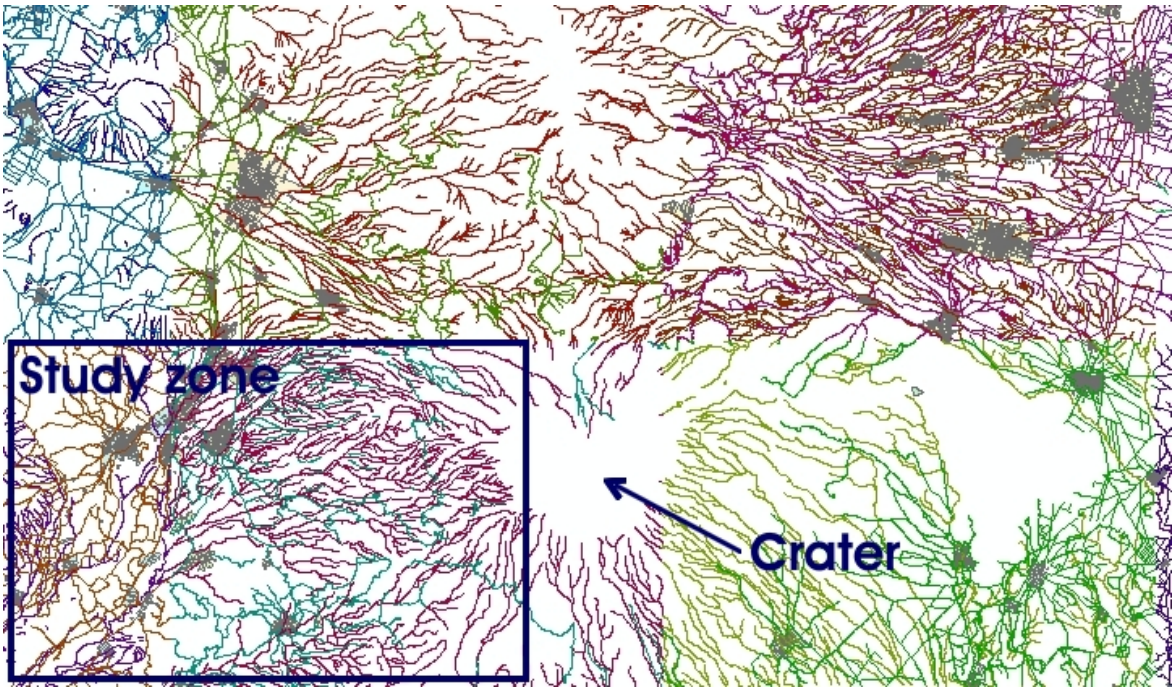


Figure 6.14. Popocatépetl volcano.

The experiments will focus on identifying relationships and characteristics shared among settlements, roads and rivers in the study zone. Suppose we want to know characteristics shared by these elements that can help us to implement/evaluate evacuation plans in case of a volcanic contingence. For example, characteristics of roads starting in or crossing a settlement, material used to build those roads and their current status (i.e., paved, unpaved), characteristics of the roads and rivers meeting a relationship (i.e., they cross, touch) in the zone, rivers near to settlements that in case of huge pluvial concentration might represent a potential risk.

6.2.2. Use-case: Popocatépetl

To illustrate the capabilities of our model for modeling and mining spatial data we will use as our study zone that shown in Figure 6.15 (Southwest side of the volcano crater). The experiment is focused on discovering characteristics among roads, rivers, and settlements in the zone. The presentation of the generated results will be structured in the following way: We will present discovered patterns among roads and rivers, roads and settlements, and rivers and settlements in the study zone. We chose this structure because we wanted to illustrate the user's capabilities to organize the data in order to create different sceneries to represent and mine spatial data.

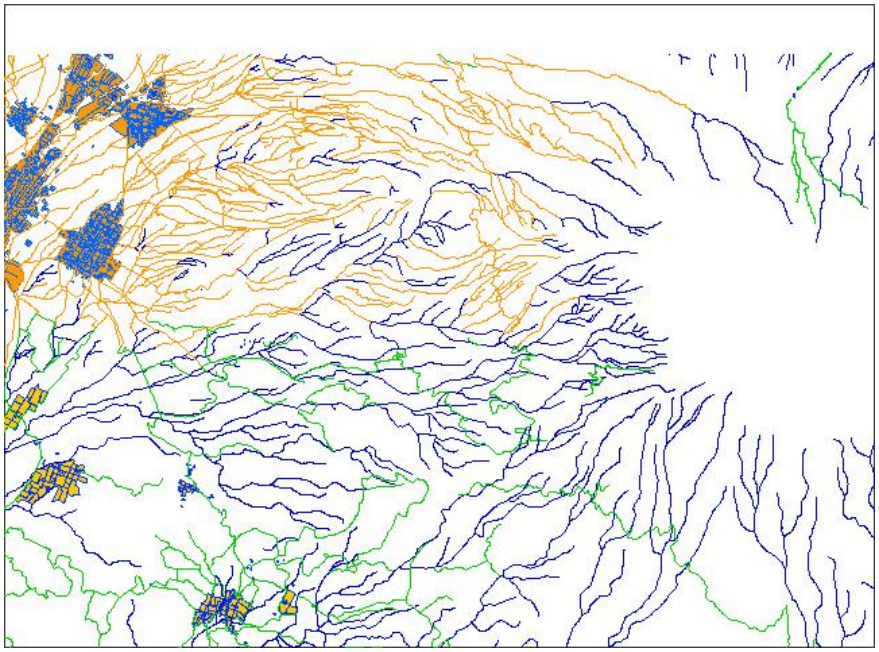


Figure 6.15. Popocatépetl volcano: study zone.

Therefore, we used our prototype system to select the river, settlement, and road spatial layers. The next step was to select the spatial relationships to be validated among the spatial

objects under consideration. To develop our test, we took advantage of a special feature implemented in the Spatial Oracle module (our SDBMS system): Oracle has the capability to examine two geometry objects to determine their topological spatial relationship. Moreover, it is possible to indicate that the same SDBMS determines and returns the topological relationship that best matches the geometries.

So, to create our dataset, we first selected the spatial layers to work with and then we evaluated the relationships among the spatial objects contained in the spatial layers. The experiments were implemented based on the following parameters:

- Spatial layers: *rivers*, *settlements* and *roads*.
- Spatial relation: topological relations supported by Oracle Spatial.
- Spatial graph-based model: all models.

The experiment was performed using the five proposed graph-based models. The objective was to evaluate the results generated by each of them. Additionally, we ran the test using no overlap, standard overlap and limited overlap features. In the following figures we show the generated results in the experiment. In the figures the generated result by using no overlap is labeled as "*a)*", via standard overlap is labeled as "*b)*", and through limited overlap is labeled as "*c)*". In the case of limited overlap, we told Subdue to allow overlap only for vertices representing roads in the zone because this element represents a primary item in our study domain (evaluation and implementation of population evacuation plans).

Since our intention in the experiment was to compare the generated results using the proposed graph-based models, we selected (and reported) as the most significant discovered pattern (by using no overlap, standard overlap and limited overlap), the one covering the following restrictions:

- Complete pattern. A pattern reporting at least two spatial objects (i.e., road and river), the spatial relation among them (i.e., touch), and some non-spatial attribute(s) (i.e., "road category unpaved" and "river category draining").
- Maximum number of reported instances. A pattern with the highest score of reported instances of a substructure.

The following subsections present the generated results using model 1 to 5. By each model we describe the discovered patterns according to the proposed structure to mine data: road and river, road and settlement, and finally river and settlement. At the end of Section 6.2.2, we present comparison tables and conclusions of the generated results by each model.

6.2.2.1 Model #1 - base model

Road and River. Figure 6.16 shows the most significant discovered pattern between roads and rivers. The pattern describes a relationship among "road category unpaved overlapping a river category draining" in the zone. This pattern may be considered as an indicator of the number of roads that need to be supervised in case of a volcanic contingence since the material type they are built with, and because they cross rivers (the lecture may be done in inverse order) that in case of huge pluvial concentration may overflow and make roads useless. Subdue found by using no overlap 46 instances (in the second iteration) of the

pattern; via standard overlap it found 85 instances (in the first iteration), and through limited overlap it also found 85 instances (in the second iteration). As we can see in the figure, standard and limited overlap found the same number of instances of the pattern, but limited overlap required two iterations to find the same pattern. However, this fact does not mean that standard overlap is better than limited overlap because analyzing the overall processing time required by limited overlap to finish the substructure discovery phase we note that it was lower than the required by standard overlap.

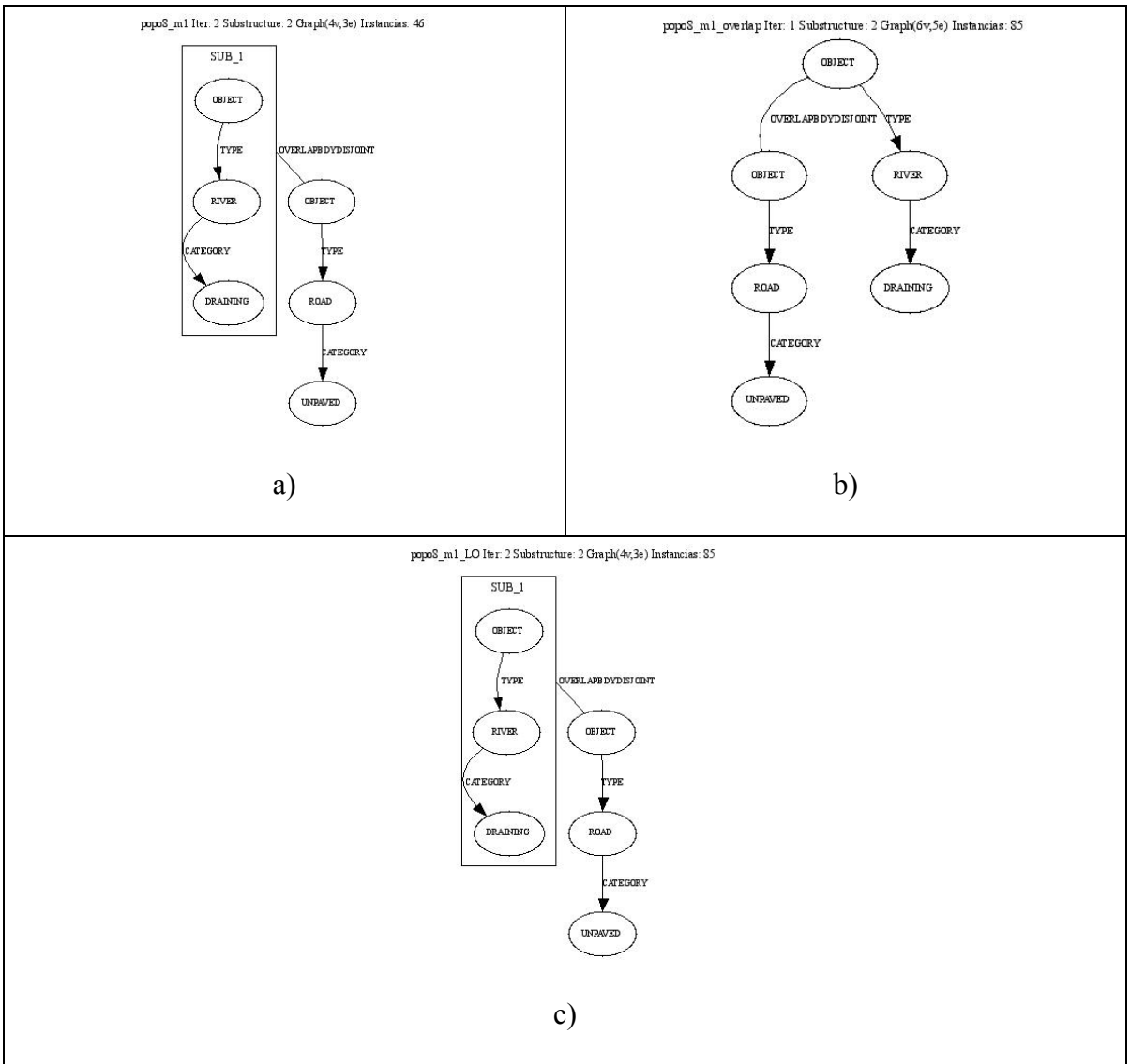


Figure 6.16. Relationships among roads and rivers by using model #1.

Road and Settlement. The most significant discovered pattern found in this experiment describes a relationship among "road category unpaved touching a settlement category construction" in the zone as shown in Figure 6.17. "Settlement category construction" represents in the Popocatépetl's settlement spatial layer inhabit areas with huge population, buildings and several constructions used to offer services to people. If we assume that

people may require to be evacuated in case of an eruption and that the roads that will be used are unpaved then this situation may become a problem (i.e., a bottleneck, water and soil may become mud and this may make roads useless). For this experiment Subdue found via no overlap 6 instances (in the ninth iteration) of the pattern, through standard overlap it found 9 instances (in the fourth iteration), and by using limited overlap it discovered 8 instances (in the tenth iteration). In all cases Subdue was able to discover the same pattern; the difference was the number of computed iterations required to discover it.

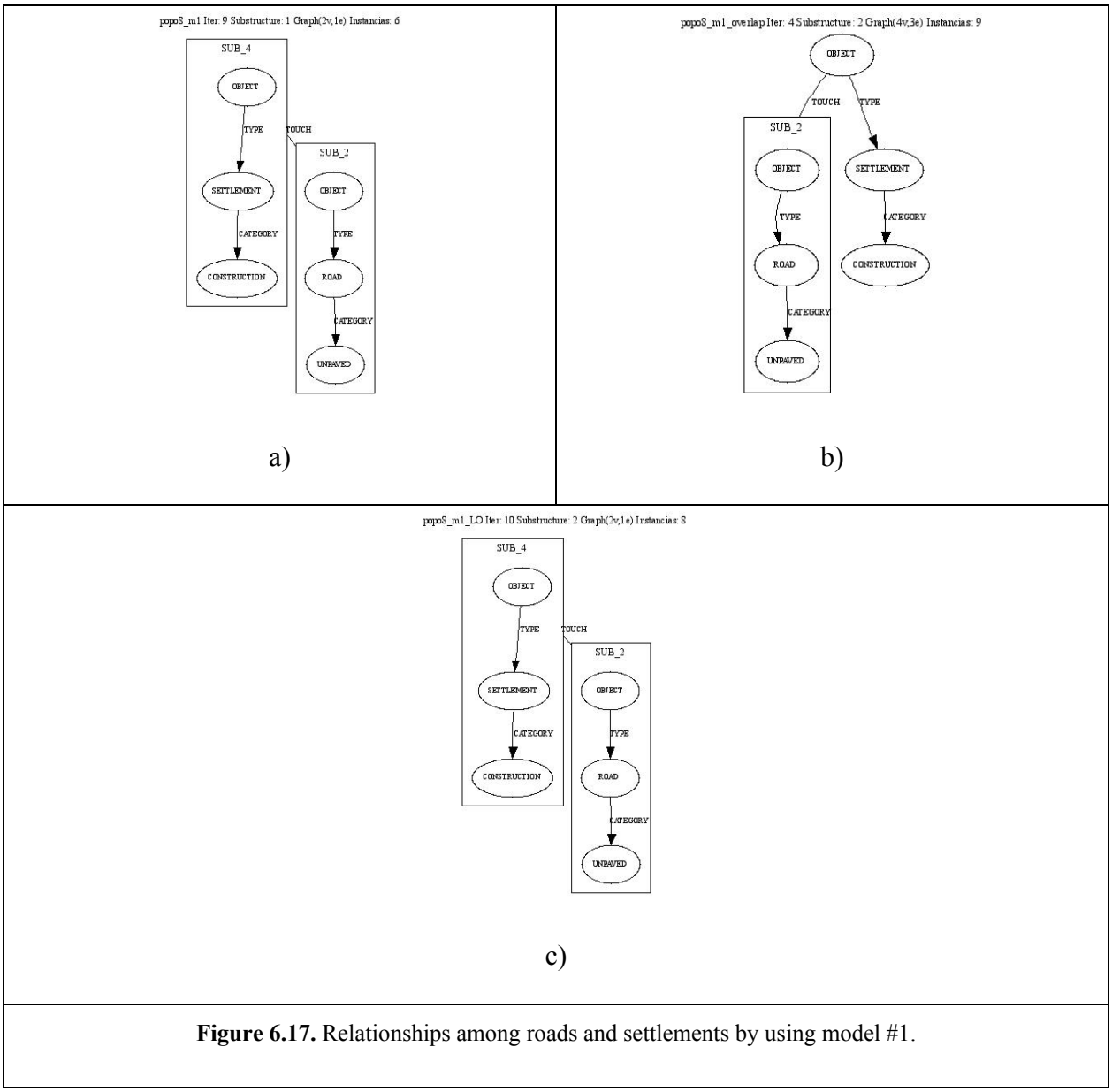


Figure 6.17. Relationships among roads and settlements by using model #1.

River and Settlement. Figure 6.18 shows the most significant discovered pattern for these spatial objects. It describes a relationship among "river category draining crossing a settlement category either (a) block or (b)(c) construction" in the zone. "Settlement category block" represents in the Popocatépetl´s settlement spatial layer inhabit areas but with small population, in fact there exist several uninhabited areas, few buildings and

constructions. The pattern may be used to identify potential flooding zones because it represents rivers close to (may be some of them crossing) areas inhabited by people. Through no overlap Subdue discovered 5 instances (in the twelfth iteration) of the pattern, by using standard overlap it also found 5 instances (in the eighth iteration), and finally, via limited overlap it also found 5 instances (in the eighth iteration). For this experiment Subdue discovered the same pattern in the three cases, however, by using standard overlap and limited overlap the understanding of the pattern is simpler.

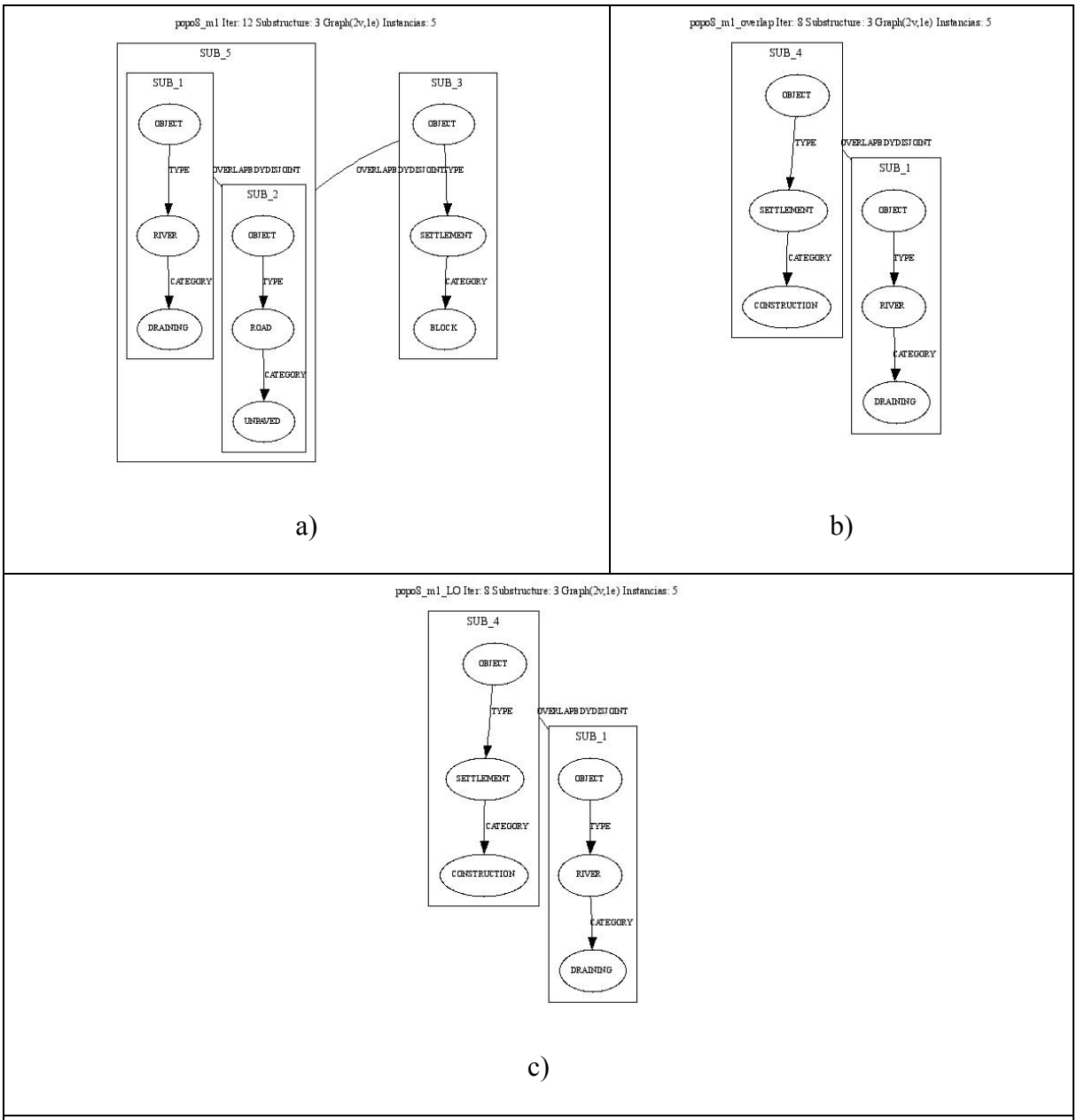


Figure 6.18. Relationships among rivers and settlements by using model #1.

The previous experiment was done using model 1, now we perform the same experiment using model 2 to 5. We will be focused to compare the same discovered patterns for the three "object-object" structures (i.e., road-river, road-settlement, and river-settlement). The

generated results are shown in the figures following the same report schema: those using no overlap are labeled as "*a)*", those using standard overlap are labeled as "*b)*", and finally those created with limited overlap are labeled as "*c)*". The most significant differences between the generated results are the number of reported instances and the number of iterations needed to discover the pattern. More iterations means more processing time to discover the pattern.

6.2.2.2 Model #2 - single replication of relation types, complete information

Road and River. By means of model #2 Subdue discovered the pattern (among road and river) shown in Figure 6.19. Subdue found by using no overlap 41 instances (in the second iteration) of the pattern, via standard overlap it found 85 (in the third iteration), and through limited overlap it discovered 64 (in the second iteration). All cases reported "road category unpaved and river category draining" as the predominant spatial objects.

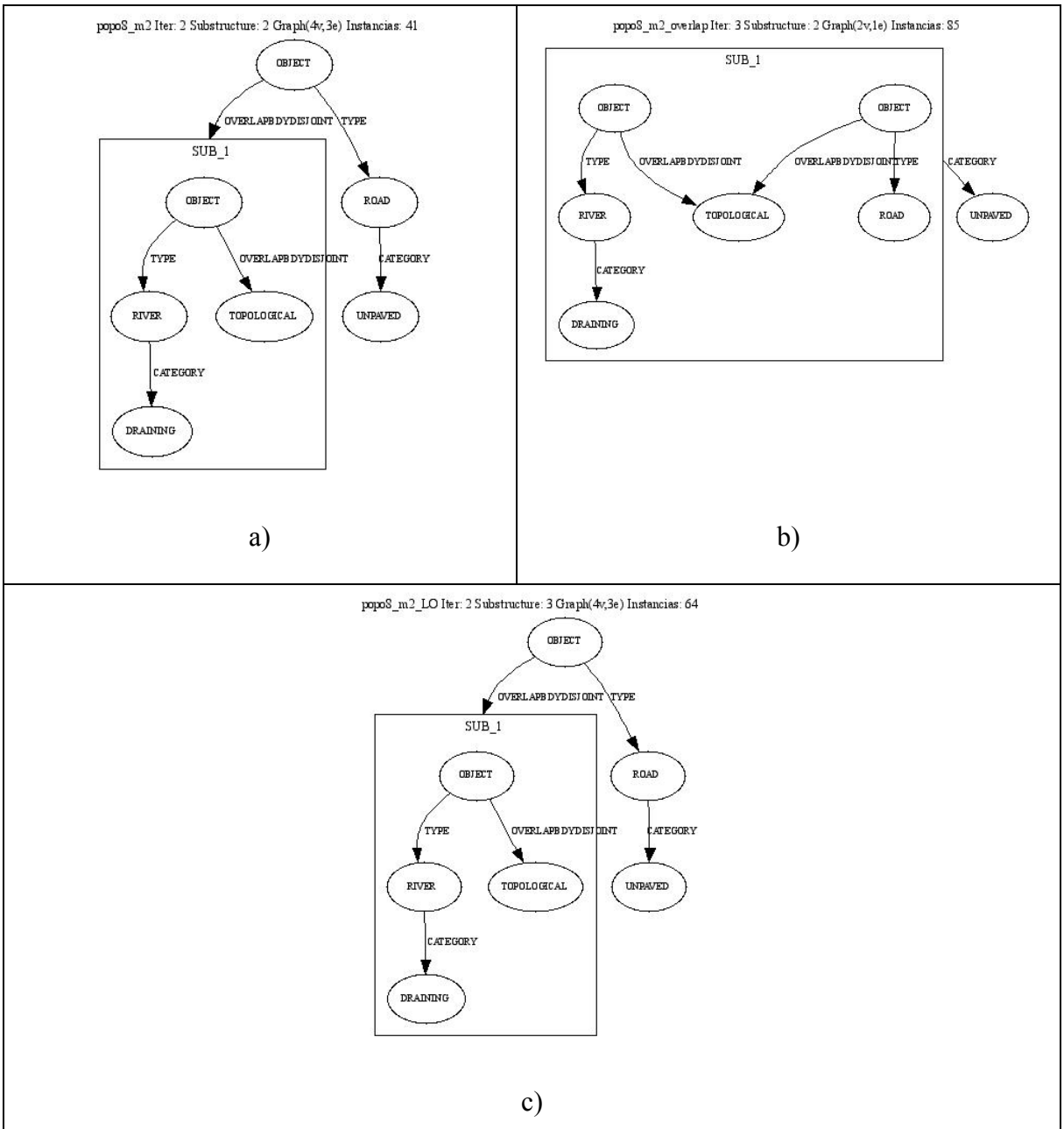


Figure 6.19. Relationships among roads and rivers by using model #2.

Road and Settlement. For these spatial objects Subdue was able to discover, by means of model #2, the pattern shown in Figure 6.20. Via no overlap Subdue found 5 instances (in

the fourteenth iteration) of the pattern, through standard overlap it discovered 8 (in the sixth iteration), and by using limited overlap it found 7 (in the tenth iteration). No overlap reported "settlement category block" whereas standard and limited overlap reported more instances of "settlement category construction".

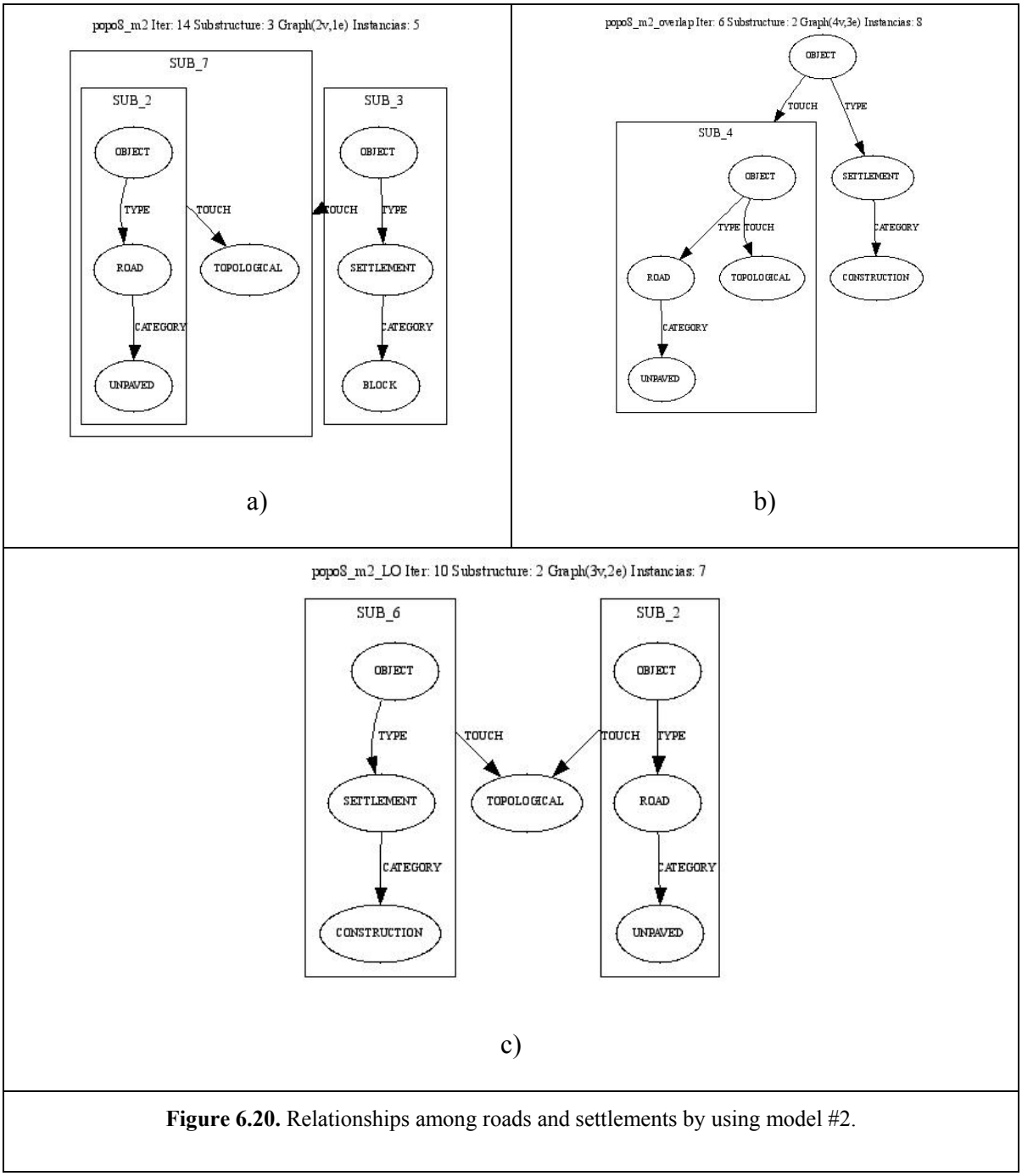


Figure 6.20. Relationships among roads and settlements by using model #2.

River and Settlement. The pattern discovered by means of model #2 among these objects is shown in Figure 6.21. Through no overlap Subdue found 5 instances (in the sixteenth iteration) of the pattern, by using standard overlap it discovered 10 (in the seventh iteration), and via limited overlap it found 5 (in the fourteenth iteration). No overlap and limited overlap reported "settlement category construction" whereas limited overlap reported more instances of "settlement category block".

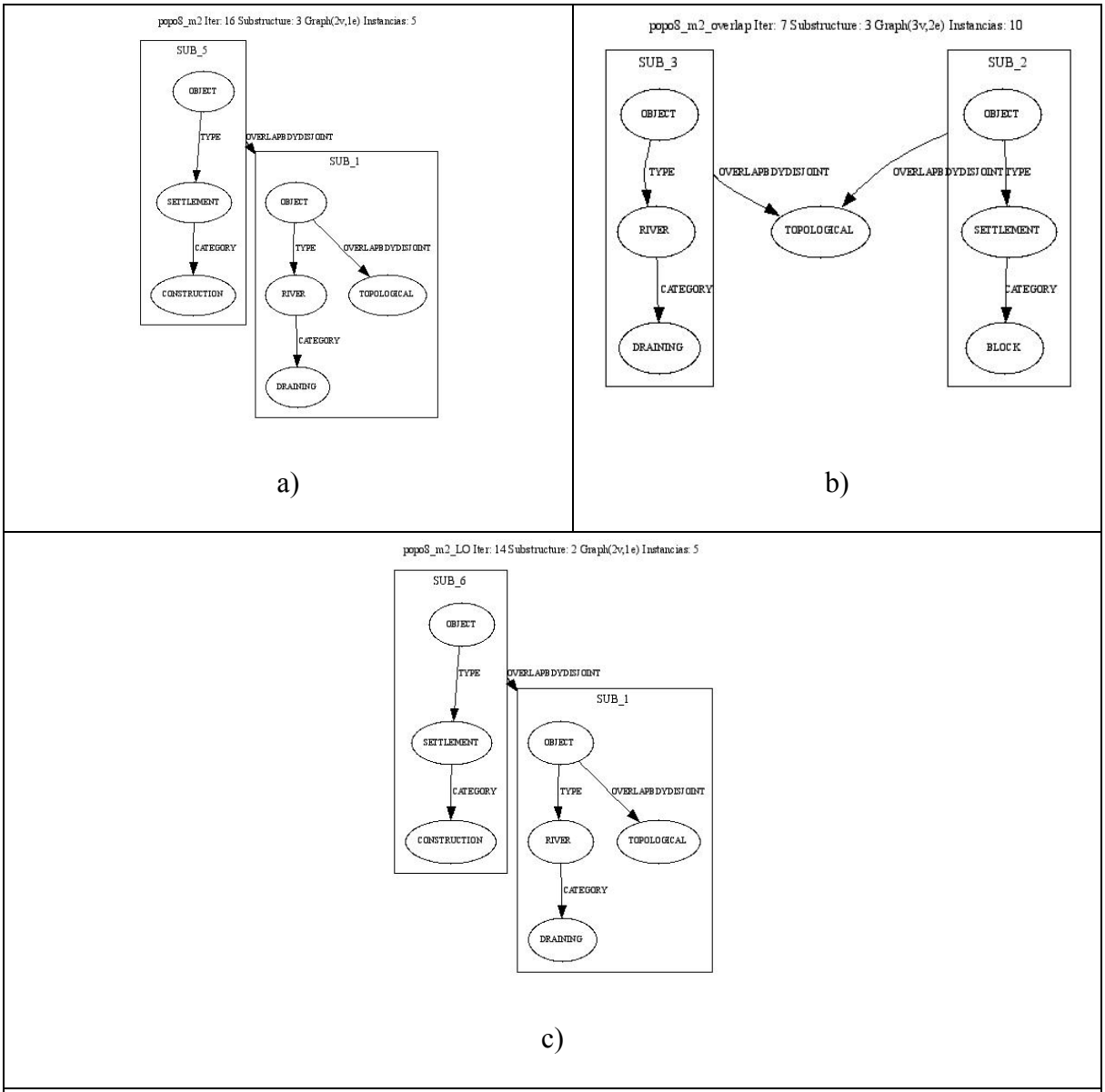


Figure 6.21. Relationships among rivers and settlements by using model #2.

6.2.2.3 Model #3 - double replication of relation types, no complete information

Road and River. The pattern discovered, by means of model #3, for these objects is shown in Figure 6.22. Through no overlap Subdue found 39 instances (in the second iteration) of the pattern, by using standard overlap it found 85 (in the second iteration), and via limited overlap it discovered 34 (in the ninth iteration). In all cases Subdue reported as the predominant spatial objects “road category unpaved and river category draining”.

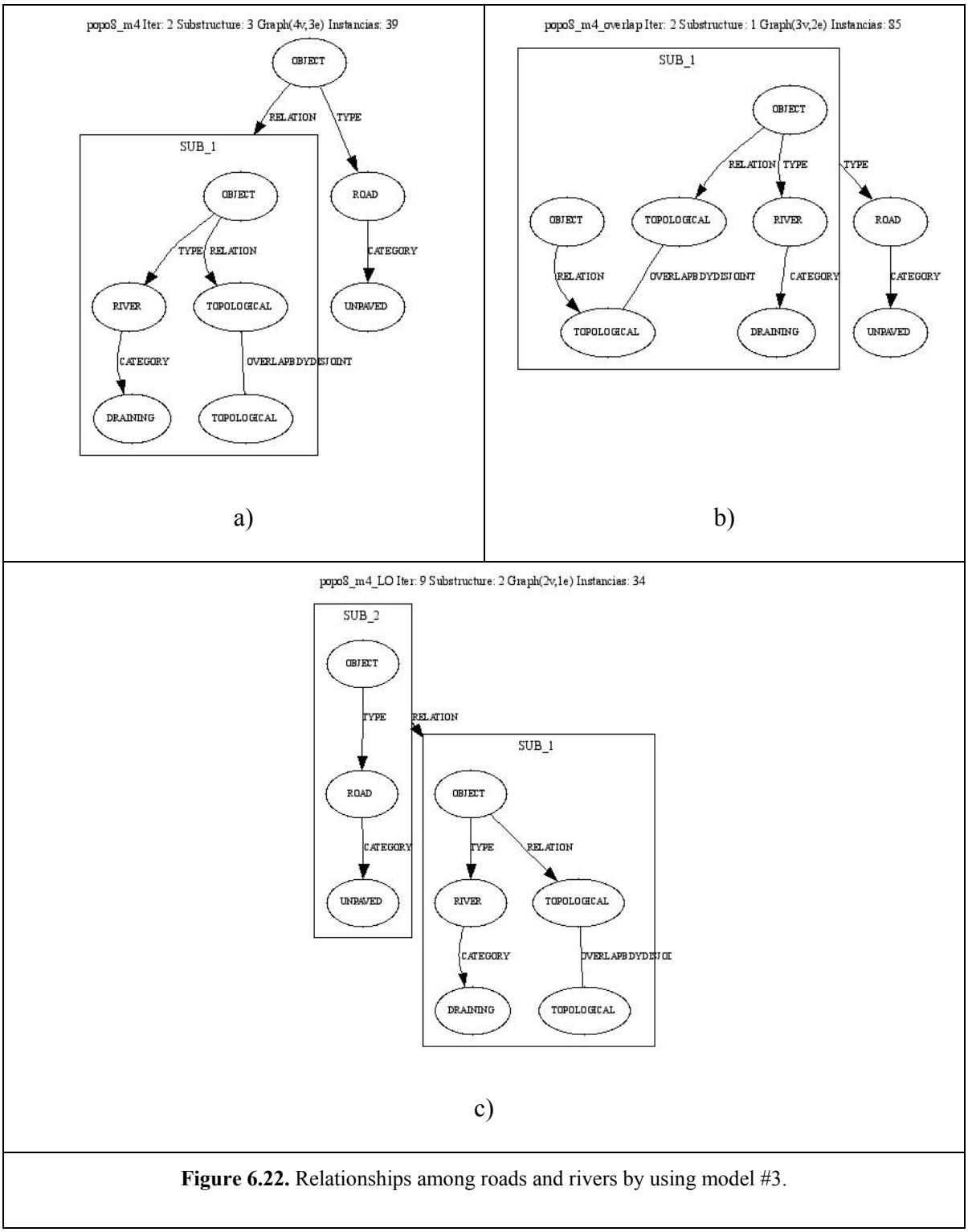


Figure 6.22. Relationships among roads and rivers by using model #3.

Road and Settlement. By means of model #3 Subdue discovered the pattern shown in Figure 6.23. Subdue found by using no overlap 4 instances (in the fifteenth iteration) of the pattern, via standard overlap it found 8 (in the sixth iteration), and through limited overlap it discovered 5 (in the thirteenth iteration). No overlap and limited overlap reported "settlement category block" as the predominant spatial object whereas standard overlap reported "settlement category construction".

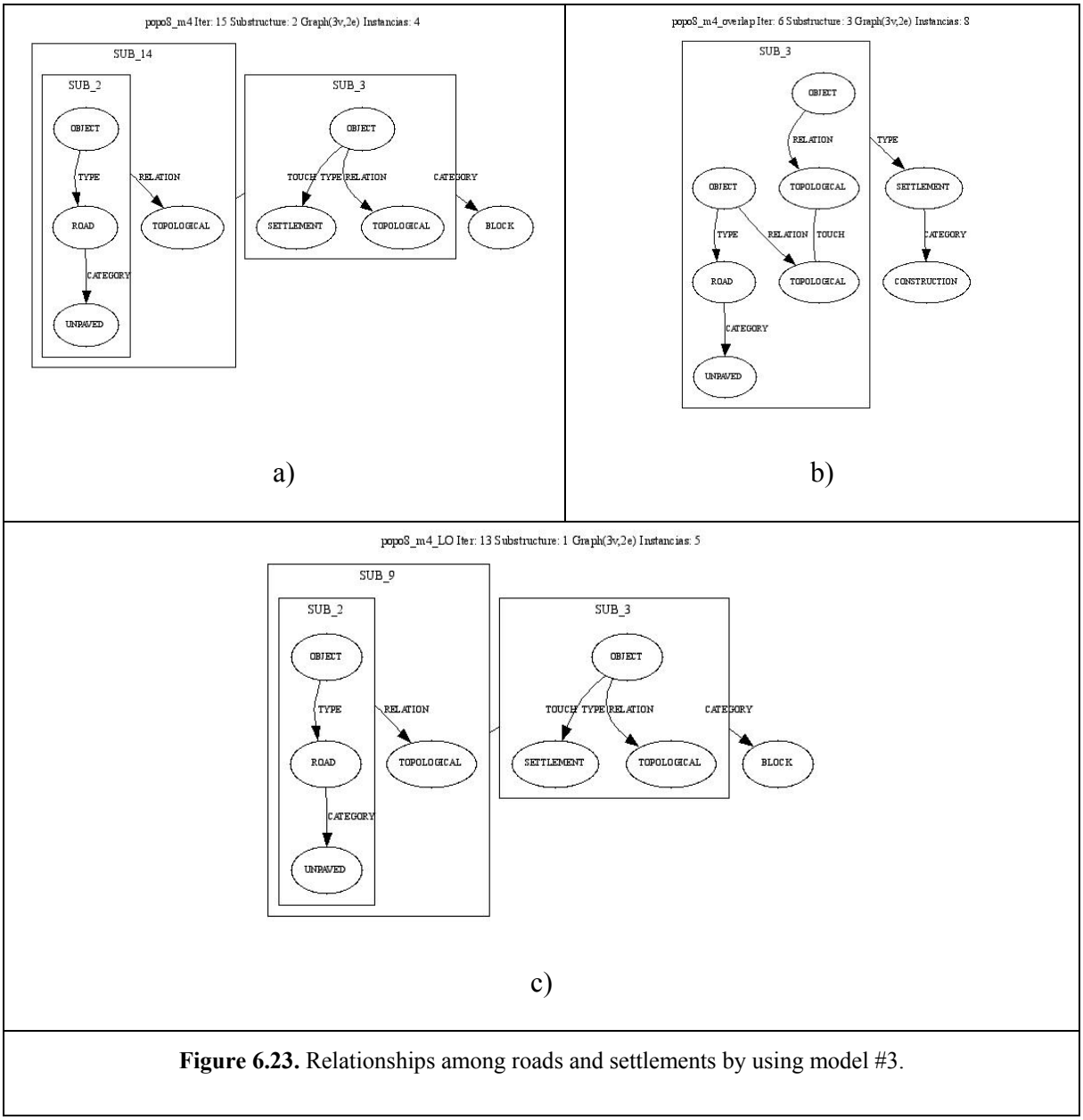


Figure 6.23. Relationships among roads and settlements by using model #3.

River and Settlement. For these spatial objects Subdue was able to discover, by means of model #3, the pattern shown in Figure 6.24. Via no overlap Subdue found 5 instances (in the twelfth iteration) of the pattern, through standard overlap it found 19 (in the fourth iteration), and by using limited overlap it found 5 (in the tenth iteration). No overlap and limited overlap reported less instances of "settlement category construction" whereas standard overlap reported more instances of "settlement category block".

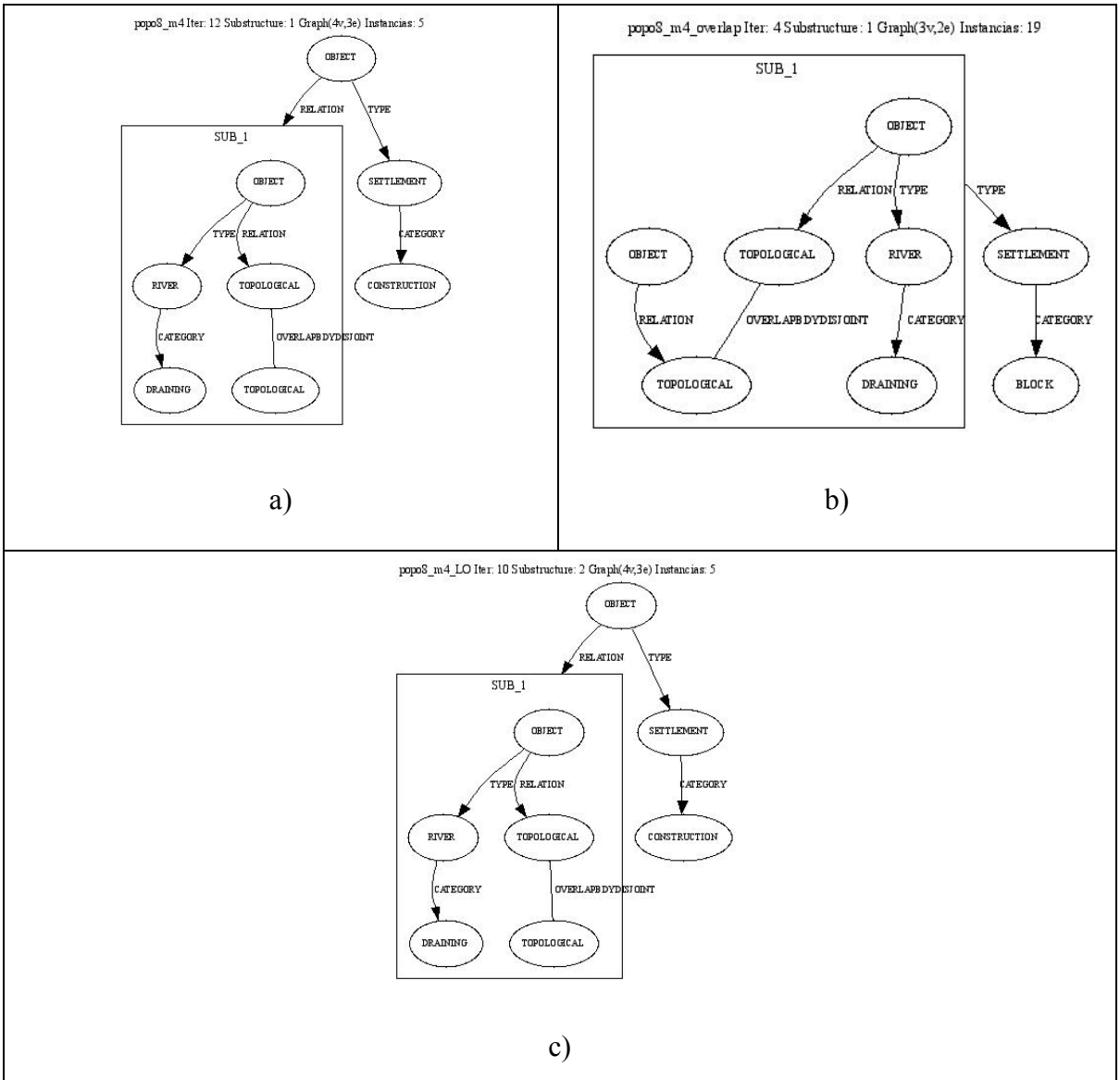


Figure 6.24. Relationships among rivers and settlements by using model #3.

6.2.2.4 Model #4 - single replication of relation types, no complete information

Road and River. For these spatial objects Subdue was able to discover, by means of model #4, the pattern shown in Figure 6.25. Via no overlap Subdue found 39 instances (in the second iteration) of the pattern, through standard overlap it discovered 85 (in the first iteration), and by using limited overlap it found 60 (in the second iteration). All cases reported as the predominant spatial objects "road category unpaved and river category draining".

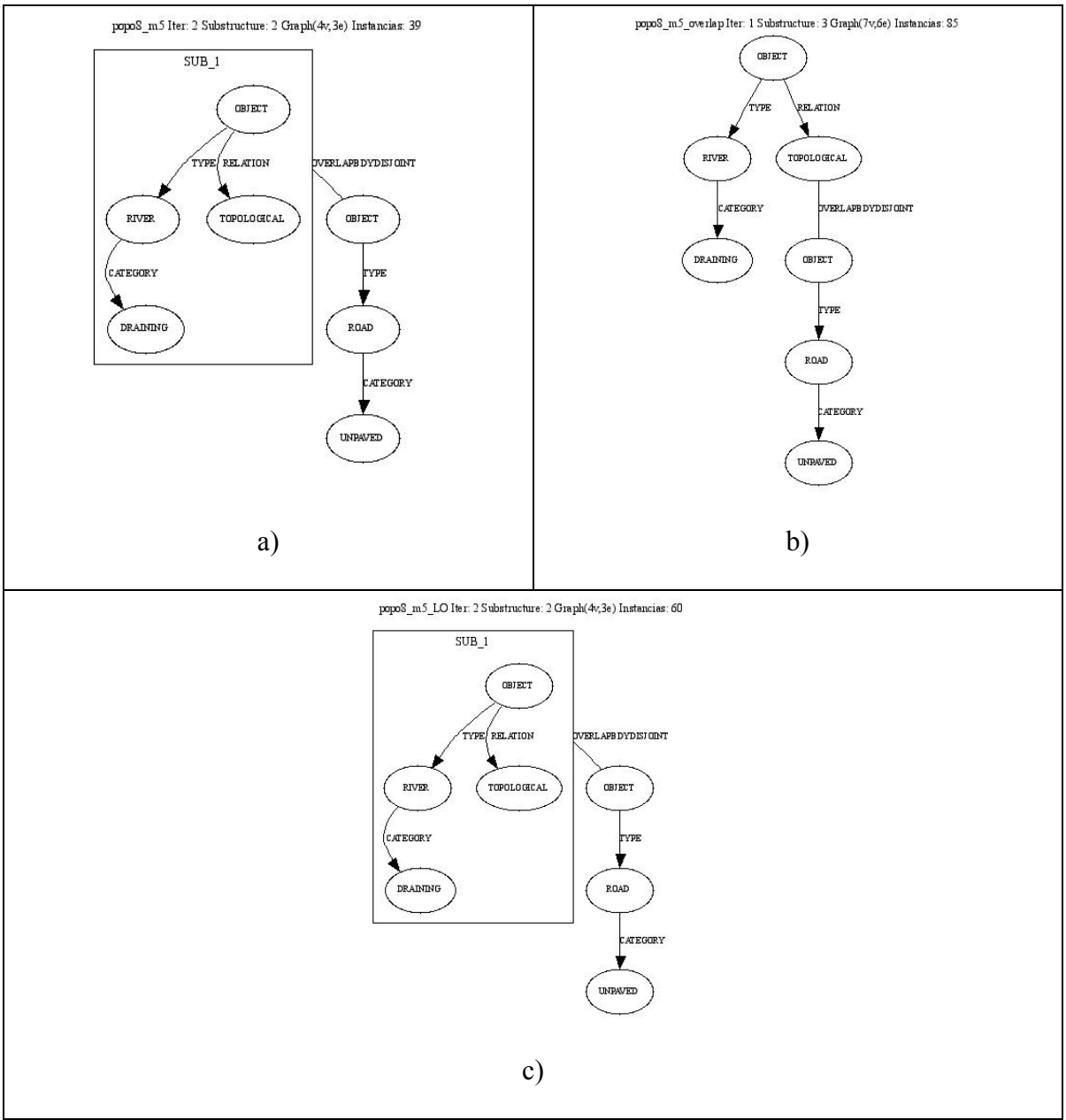


Figure 6.25. Relationships among roads and rivers by using model #4.

Road and Settlement. The pattern discovered, by means of model #4, for these objects is shown in Figure 6.26. Through no overlap Subdue discovered 6 instances (in the twelfth iteration) of the pattern, by using standard overlap it found 8 (in the tenth iteration), and via

limited overlap it found 8 (in the seventh iteration). The representative spatial objects are "road category unpaved and settlement category construction" in all cases.

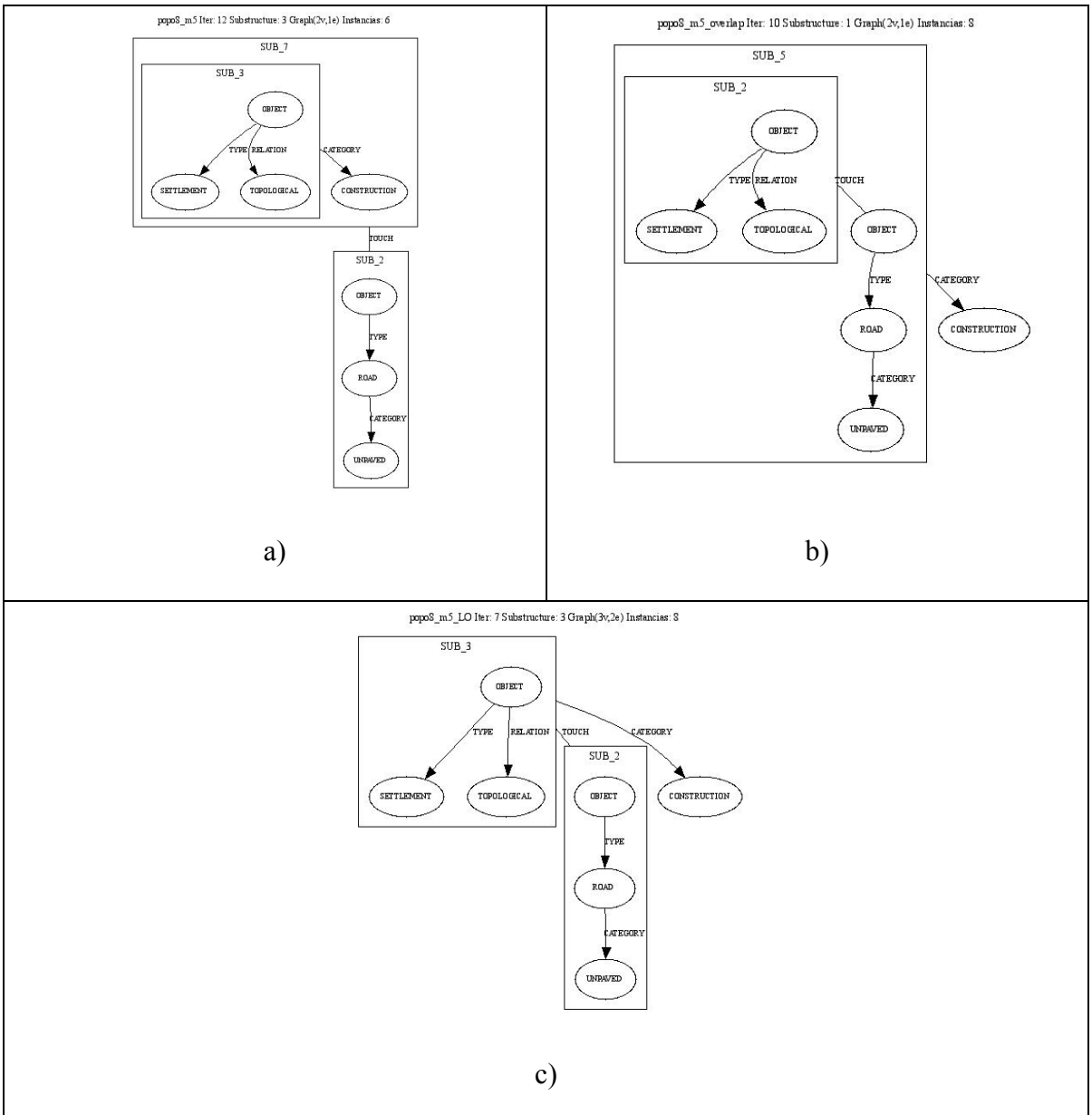


Figure 6.26. Relationships among roads and settlements by using model #4.

River and Settlement. By means of model #4 Subdue discovered the pattern shown in Figure 6.27. Subdue found by using no overlap 5 instances (in the sixth iteration) of the

pattern, via standard overlap it found 10 (in the sixth iteration), and through limited overlap it found 5 (in the eleventh iteration). No overlap and limited overlap reported less instances of "settlement category construction" whereas standard overlap reported more instances of "settlement category block".

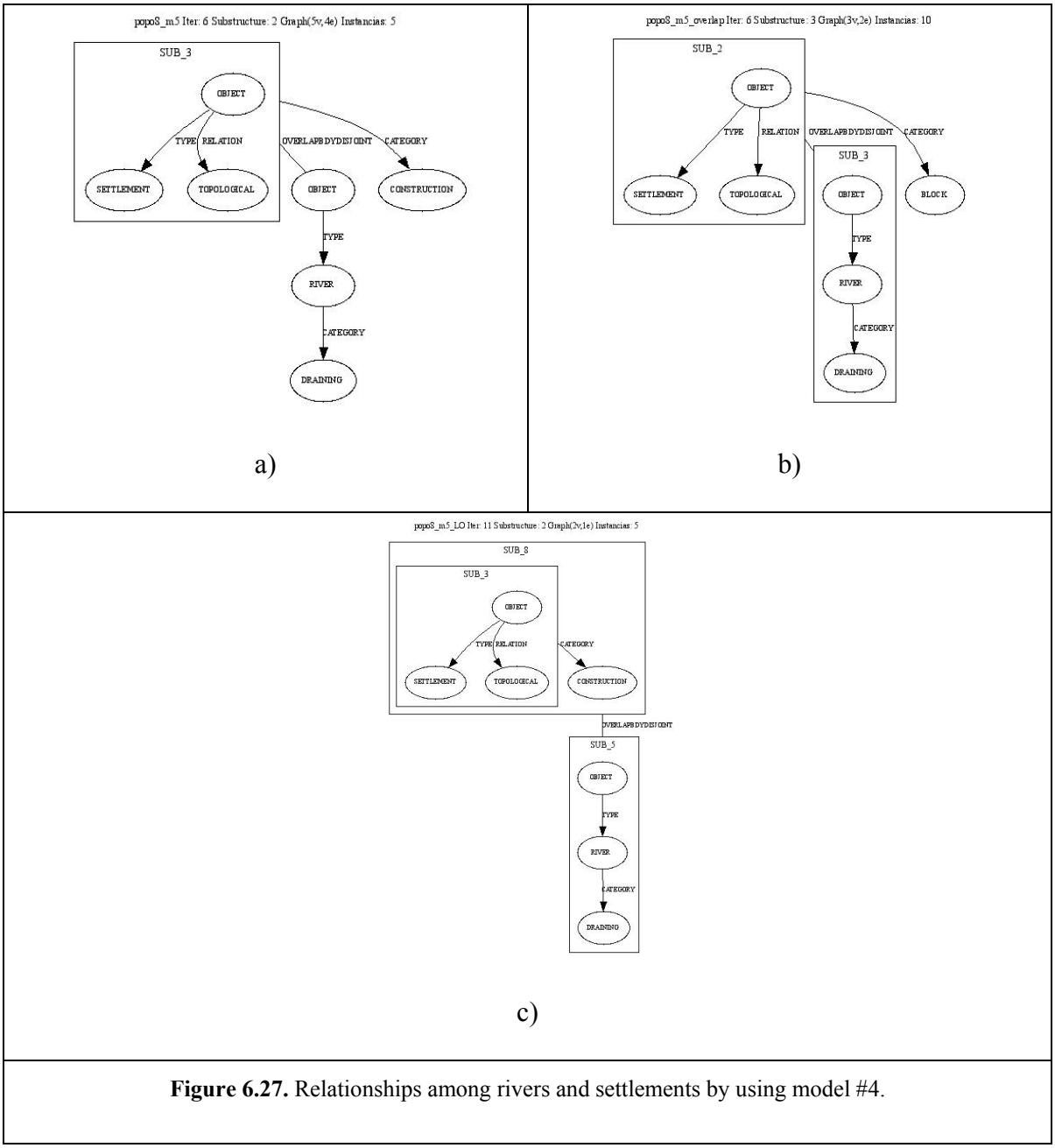
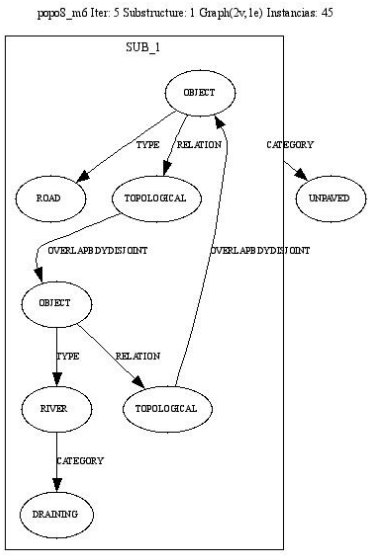


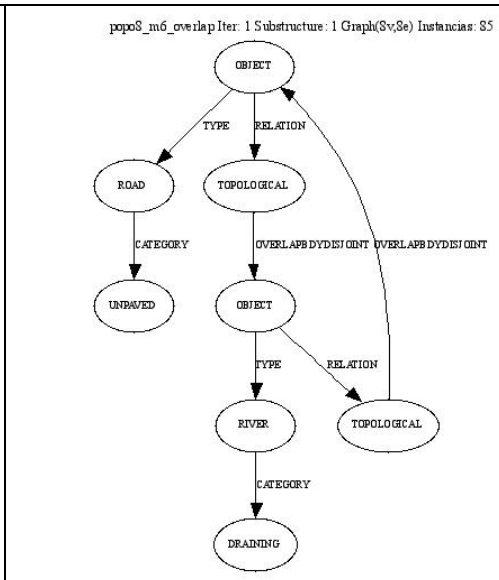
Figure 6.27. Relationships among rivers and settlements by using model #4.

6.2.2.5 Model #5 - double replication of relation types, complete information

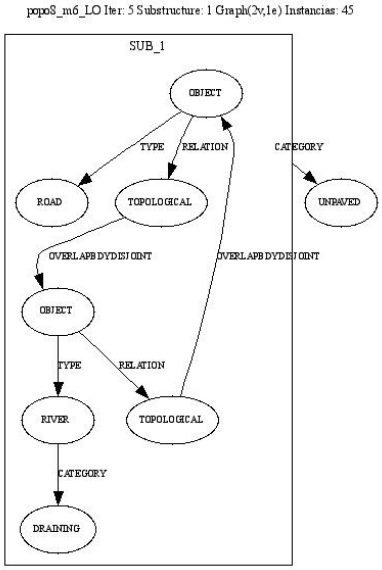
Road and River. The pattern discovered, by means of model #5, for these objects is shown in Figure 6.28. Through no overlap Subdue discovered 45 instances (in the fifth iteration) of the pattern, by using standard overlap it found 85 (in the first iteration), and via limited overlap it found 45 (in the fifth iteration). Subdue reported in all cases as the representative spatial objects “road category unpaved and river category draining”.



a)



b)



c)

Figure 6.28. Relationships among roads and rivers by using model #5.

Road and Settlement. By means of model #5 Subdue discovered the pattern shown in Figure 6.29. Subdue found by using no overlap 6 instances (in the seventh iteration) of the pattern, via standard overlap Subdue could not find a complete pattern (in the figure the category of the settlement is not reported), and through limited overlap it found 7 (in the seventh iteration). No overlap and limited overlap reported as the representative spatial objects "road category unpaved and settlement category construction".

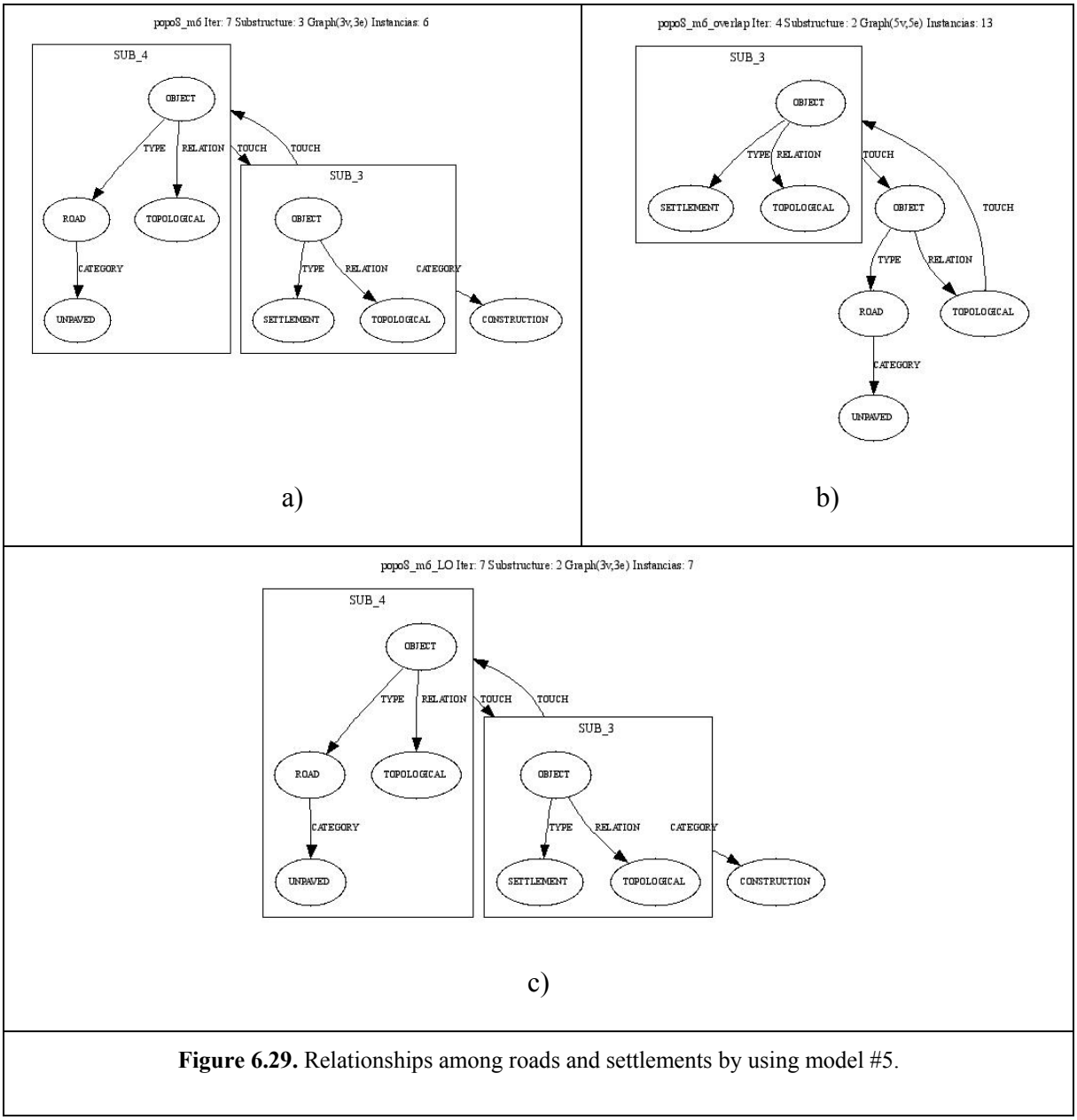


Figure 6.29. Relationships among roads and settlements by using model #5.

River and Settlement. For these spatial objects Subdue was able to discover, by means of model #5, the pattern shown in Figure 6.30. Via no overlap Subdue discovered 5 instances (in the thirteenth iteration) of the pattern, through standard overlap it found 5 (in the sixth iteration), and by using limited overlap it also found 5 (in the tenth iteration). Subdue reported in all cases "river category draining and settlement category construction" as the predominant spatial objects.

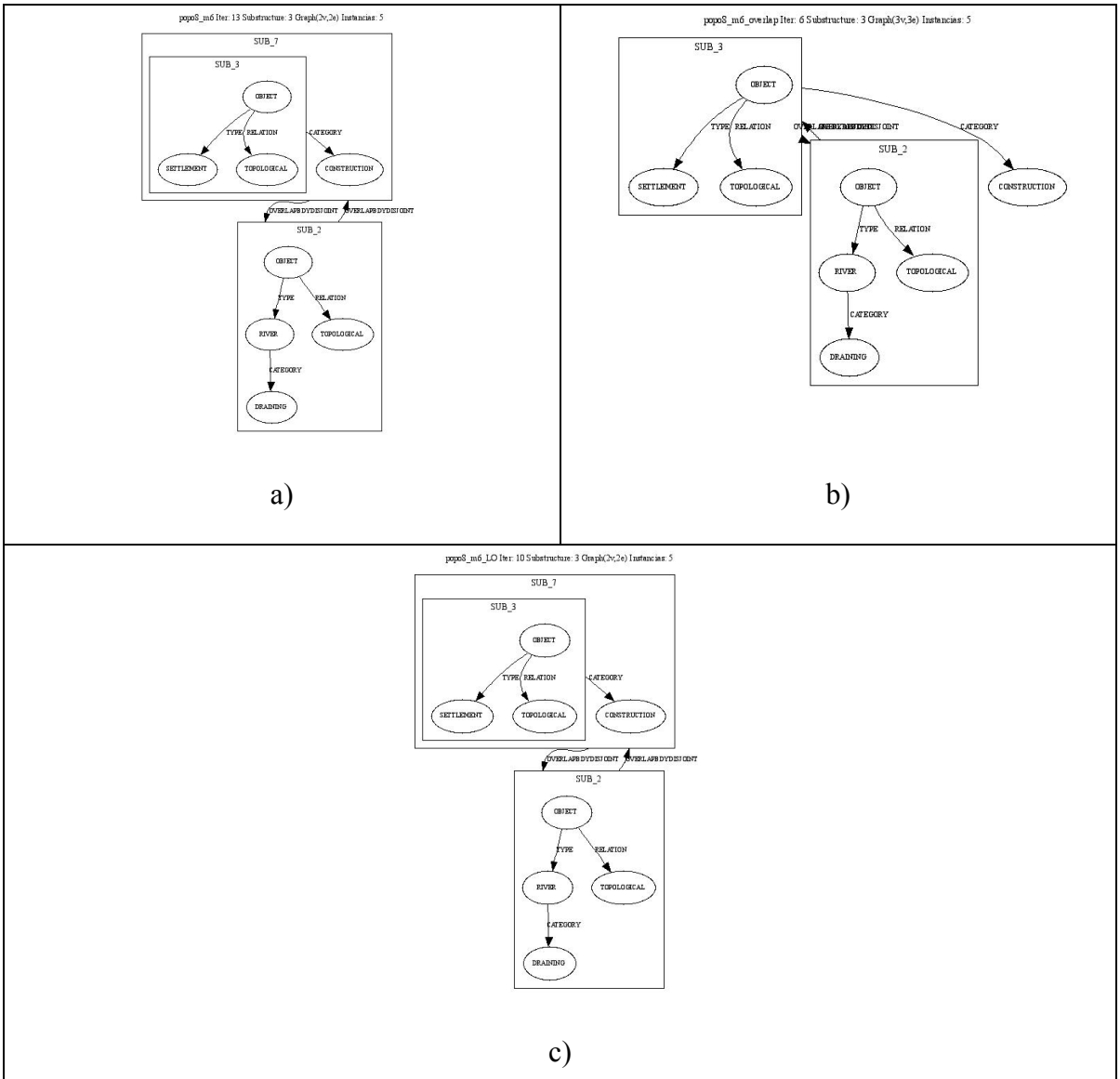


Figure 6.30. Relationships among rivers and settlements by using model #5.

Table 6.1 presents a comparison, by each model and overlap feature, among the number of discovered instances (patterns) and the number of iterations needs to discover them. For example, by using model #1, Subdue found via no overlap 46 instances (in the second iteration) of a "complete" pattern (according to our definition for reporting a complete pattern) involving the spatial objects road-river. A higher score means a model allowing us to discover more instances of a substructure. Remember Subdue reported as the best pattern (by iteration) a substructure with the highest score of discovered instances of that substructure. This comparison is reported by each "object-object" structure (i.e., road-river).

Note: NO (No overlap), SO (Standard overlap), LO (Limited overlap).

	Model #1			**Model #2**			**Model #3**			**Model #4**			**Model #5**		
	NO	**SO**	**LO**	**NO**	**SO**	**LO**	**NO**	**SO**	**LO**	**NO**	**SO**	**LO**	**NO**	**SO**	**LO**
Road-River															
Instances	46	85	85	41	85	64	39	85	34	39	85	60	45	85	45
Iterations	2	1	2	2	3	2	2	2	9	2	1	2	5	1	5
Road-Settlement															
Instances	6	9	8	5	8	7	4	8	5	6	8	8	6	0	7
Iterations	9	4	10	14	6	10	15	6	13	12	10	7	7	0	7
River-Settlement															
Instances	5	5	5	5	10	5	5	19	5	5	10	5	5	5	5
Iterations	12	8	8	16	7	14	12	4	10	6	6	11	13	6	10

Table 6.1. Instances/iterations by each graph-based model: Popocatépetl use-case.

From the previous table we can see that Subdue (setting no overlap -NO- in all cases) reported using model #1, 46 instances in the second iteration, of a complete pattern among the road and river spatial objects. Using model #2, Subdue reported 41 instances of a complete pattern in the second iteration. Using model #3, Subdue reported 30 instances of a complete pattern in the second iteration and so on. Finally, we can identify that model #1 was the best model to find the highest number of a complete pattern among these spatial objects (shown in pink color in Table 6.2).

	Model #1	**Model #2**	**Model #3**	**Model #4**	**Model #5**
	NO	**NO**	**NO**	**NO**	**NO**
Road-River					
Instances	46	41	39	39	45
Iterations	2	2	2	2	5

Table 6.2. The best model to discover complete patterns among the road and river spatial objects (no overlap)

From Table 6.1 we created Table 6.3. This table presents a comparison of maximum/minimum discovered instances by each overlap feature. A model with the highest score is better since it allows discovering more instances of a substructure (patterns). The comparison is presented for each "object-object" structure (i.e., road-river).

For example, we have already mentioned that model #1 was the best model to find complete patterns (setting no overlap) among the road and river spatial objects. Subdue reported 46 discovered instances of a complete pattern in the second iteration (the highest score). See Table 6.1 for details.

On the other hand, model #3 and model #4 were the worst models to find complete patterns among the road and river spatial objects. For example, using model #3, Subdue only reported 39 instances of a complete pattern in the second iteration.

	Maximum	Minimum
Road-River		
No overlap	model #1 (second iteration)	models #3 and #4 (second iteration)
Standard overlap	models #1, #4 and #5 (first iteration)	model #2 (third iteration)
Limited overlap	model #1 (second iteration)	model #3 (ninth iteration)
Road-Settlement		
No overlap	model #5 (seventh iteration)	model #3 (fifteenth iteration)
Standard overlap	model #1 (fourth iteration)	model #5 (no complete pattern)
Limited overlap	model #4 (seventh iteration)	model #3 (thirteenth iteration)
River-Settlement		
No overlap	model #4 (sixth iteration)	model #2 (sixteenth iteration).
Standard overlap	model #3 (fourth iteration)	model #1 (eighth iteration).
Limited overlap	model #1 (eighth iteration)	model #2 (fourteenth iteration)

Table 6.3. Max/Min of discovered instances by "object-object"/overlap feature.

From Table 6.3 we can identify that model #1 was the best model to discover complete patterns in most of the cases. The second best most was model #4. On the other hand, the worst model to discover complete patterns was model #3.

Table 6.4 presents a comparison among the average of discovered instances by model. Higher score means a model allowing us to discover more instances of a substructure (patterns). Each value represents the average of discovered substructures by using no

overlap, standard overlap and limited overlap features. For example, the value 72.0 in column "Model #1" and row "Road-River" is the average computed from the values 46, 85 and 85 obtained from Table 6.1. The comparison is reported by each "object-object" structure (i.e., road-river). We can see in the table that model #1 was the best model to find complete patterns. It was followed by model #2. The worst model was model #5.

	Model #1	**Model #2**	**Model #3**	**Model #4**	**Model #5**
Road-River	72.0	63.3	52.7	61.3	58.3
Road-Settlement	7.7	6.7	5.7	7.3	4.3
River-Settlement	5.0	6.7	9.7	6.7	5.0

Table 6.4. Average of discovered instances by model/"object-object".

Table 6.5 presents a comparison among the average of discovered instances by model. For example, the value 19.0 in column "Model #1 no overlap (NO)" is the average computed from the values 46, 6 and 5 obtained from Table 6.1. Higher score means a model allowing us to discover more instances of a substructure (patterns). The comparison is reported by each overlap feature.

Model #1			**Model #2**			**Model #3**			**Model #4**			**Model #5**		
NO	**SO**	**LO**	**NO**	**SO**	**LO**	**NO**	**SO**	**LO**	**NO**	**SO**	**LO**	**NO**	**SO**	**LO**
19.0	33.0	32.7	17.0	34.3	25.3	16.0	37.3	14.7	16.7	34.3	24.3	18.7	30.0	19.0

Table 6.5. Average of discovered instances by model/overlap feature.

Finally, Table 6.6 presents a comparison among the average of discovered instances by model. For example, the value 28.2 in column "Model #1" is the average computed from the values 19.0, 33.0 and 32.7 obtained from Table 6.5. We can see in the table that model

#1 reported the highest score of discovered instances (according to our parameters for reporting complete instances) in this illustrative Popocatépetl domain. The following ones were model #2 and model #4 respectively. The worst model to find complete patterns as proposed was model #5.

Model #1	Model #2	Model #3	Model #4	Model #5
28.2	25.6	22.7	25.1	22.6

Table 6.6. Average of discovered instances by model.

As part of our strategy for testing our graph-based methodology for modeling and mining spatial data, we proposed five graph-based operative models derived from our general schema. However, from the analysis and interpretation of the results presented in this section we can conclude that model #1 and model #2 are (following that order) the two graph-based models that produce the best results. This conclusion is based on the quantitative and qualitative analysis implemented using this test domain. Additionally, this appreciation is also supported by the results obtained from the discovered patterns in the population census from the year of 1777 in Puebla downtown test domain described in the previous section.

6.3 Conclusion

In this chapter we have presented three illustrative use-cases of our proposal for modeling and mining spatial data. The test domains were two spatial databases, the first one related to

a population census from the year of 1777 in Puebla downtown, and the second one related to the Popocatépetl volcano.

The use-cases developed using the population census database were focused on exemplifying the graph-based model to represent together spatial data, non-spatial data and spatial relations, the new limited overlap feature implemented in the Subdue system (processing time and specialized overlapping pattern oriented search), and the generated results by the mining phase. We have presented in each use-case evaluations performed by the domain expert over the discovered patterns.

The use-case developed using the Popocatépetl database was focused on the comparison/evaluation of the generated results by each proposed graph-based model. The tests were implemented upon the supposition of evacuation plans implementation in case of volcanic contingences. We presented comparison tables describing which model(s) allow(s) to discover more instances of a substructure via no overlap, standard overlap, and limited overlap features. Subdue reported as the best pattern (by iteration) a substructure with the highest score of discovered instances of that substructure.

Next chapter presents conclusion about our research work and final remarks.