

SIMILARITE ENTRE LES RESEAUX GEOGRAPHIQUES

Ghazal Moultazem, Sèdes Florence

IRIT, Université Paul Sabatier

Toulouse, France

Moultazem.Ghazal@irit.fr, Florence.Sedes@irit.fr

Abstract— The available amount of geographic datasets has considerably grown. These data (maps) are available in different formats, in different scales, and they are usually generated by different procedures. Distinct maps that represent the same or overlapping areas (multiresolution maps) can differ both in accuracy and resolution. Therefore, an important issue is to determine whether two maps are consistent, i.e., do they represent the same area without contradictions, and if not, are they at least similar? In this paper we develop a method to assess the similarity over complex structured spatial objects that form networks. Existing approaches deal only with the change in accuracy and do not take into account the change in resolution; as a result the two maps in question must have the same number of components. In this paper we extend them to treat the change in resolution by tying each component in the map with weight according to its importance. Other improvements to existing approaches, which are based on topological properties, are proposed by considering the directional and metrical properties. This method is the first step to assess the similarity between maps with complex configurations including all the geographic features.

Keywords- geographic dataset; consistency; similarity; network.

I. INTRODUCTION

The available amount of geographic datasets is significantly growing in recent time. This is due to the increasing number of different devices collecting such data; such as remote sensing systems, environmental monitoring devices, etc. These data (maps) are available in different formats, in different scales, and they are usually generated by different procedures. Consequently distinct maps can differ both in accuracy and resolution [1]. A less precise representation means that the data contain simplification of the original representation, but the topology should not be changed. On the other hand, the reduction of the map resolution may change the topological structure of spatial object.

The necessity to deal with two or more maps in the same environment requires an effective management to these multiple representations of information. Currently, multiresolution geographic database records explicitly multiple representations covering the same geographic area at different scales. Such databases require a mechanism to detect inconsistencies among the different representations. In the future multiresolution geographic databases will be envisioned that they would derive multiple representations “on the fly” by using generalization algorithm, this will provide more flexibility to the interrogation process and decrease the amount of the data stored. Under such a scenario, quality control

mechanisms will be needed to confirm that the generalization algorithm produces consistent results. Therefore, the determination of consistency and similarity of data is an important step for both scenarios.

The assessment of consistency and similarity corresponds with comparing the data and test to see whether they contain contradiction or not. It also requires geographic domain knowledge. This paper focuses on domain knowledge about networks, which is an important data type and is shared by many geographic applications (e.g. transportation, hydrology, etc.). We extend the proposed approach which assesses the similarity between networks differing in accuracy [2], to deal with networks differing in resolutions. “Fig. 1” depicts three networks: the network (b) differs from the network (a) in accuracy because only the lengths of the segments and the angles among them have modestly been modified. On the other hand, the network (c) differs from the network (a) in resolution because some segments are dropped or aggregated.

Current similarity methods focus on the topological relations among the components and ignore metric and directional relations. This paper develops a method which takes into account these relations to improve the process similarity’s result. The method employed is based on topological relationships [3], the boundary-boundary sequence [4], the cardinal direction [5], and the approximate distance [6]. This formalization represents the first steps of research efforts for developing formal models for checking the consistency among multiple representations in heterogeneous geographic databases.

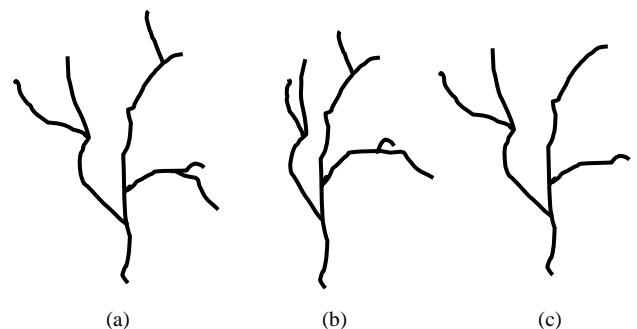


Figure 1. Difference between the change in accuracy and resolution

The remainder of this paper is organized as follows: Section 2 discusses previous works on consistency and similarity. Section 3 illustrates the data model adopted to represent the multiresolution networks. In section 4 we present our proposition to assess the similarity between networks. Finally, section 5 concludes our proposition and outlines future works.

II. RELATED WORKS

Multiresolution maps have started to emerge as a research topic in geographic information community in the 1980s [7]. It implies a considerable increase in the amount of data to be stored, introducing additional problem for the maintenance and integration of these data at different levels of detail [8]. Ideally multiple representations should be automatically derived from a single detailed representation in order to answer some specific user queries [9]. Unfortunately, this is not feasible at this time due to the inadequacy of the automated generalization procedures.

Multiresolution maps are interpreted in [10] as a set of different levels S_0, \dots, S_n where the level S_{i+1} is derived from level S_i by using some generalization operations. These operations guarantee to transform a map into another map, which is consistent with the first one with respect to topological relations. A similarity measure between maps, defined as a deviation from consistency, has also been provided in [10]. "Planar Abstract Cell Complex" has been used to define a formal model for representing multiresolution maps in [8]. In this paper the consistency test is defined at the combinatorial level by means of homeomorphisms.

A method for checking similarity between topological relations of regions is defined in [12], using the similarity values over topological relations between regions, a partial order is defined to evaluate how the relation may be changed after some generalization operations. In [13] this distance is used to create a partial order over the topological relations between lines and regions. In [12] and [13] the similarity between topological relations is computed between pairs which have the same dimension. In [14] this distance is extended to consider the changes in the dimension of the objects.

In [2], consistency among networks, defined as sets of lines (homogeneous networks) and sets of lines and regions (heterogeneous networks), is investigated. In this paper, similarity is computed between two networks which have the same number of components (small change); despite in the majority of situations the two networks don't have the same number of components. Therefore, in this paper we investigate the consistency and similarity between networks when large changes happened (some components disappeared or aggregated), by tying each component with weights representing its importance in the network and taking into account the metric and directional relation among the components.

III. NETWORK REFERENCE MODEL

A canonical representation of networks aids in the assessment of consistency and similarity. Therefore, we

introduce a set of constraint for the constituent components, which are necessary to form valid network. In this paper we address the homogenous networks, which are made up of line segments only linked by intersection points. The relations between networks and other components of the map will be addressed in a future paper.

Definition 1: A (segment) simple open line l is a continuous, non intersecting sequence of points (x,y) in \mathbb{R}^2 that can be represented by an injective continuous function as following:

$$l: [0,1] \rightarrow \mathbb{R}^2 \quad (1)$$

The segment endpoints $l(0), l(1)$ are usually referred to as the segment boundaries, whereas the other points of the line are referred to as the segment interior. "Fig. 2" shows four lines that (a-b) are valid for the segment definition whereas (c-d) are not.

The Segment may be a straight line e.g. "Fig. 2" (a), or a line that has shape points e.g. "Fig. 2" (b). To assess the similarity between two segments, we introduce the following notions: 1) the number of segment detour, 2) the detour degree of inclination. While the number segment detour is determined by counting the number of detour in the segment, the degree of inclination is estimated by computing the angle output from the detour. For computing them we joint the segment endpoints by the straight line l_s and compare its length ($leng(l_s)$) with the length of segment initial ($leng(l_0)$) According to the difference we have two cases:

1) *If the difference is less than a threshold (δ) "Fig.3" (a), the segment will be represented by the equation of the straight line which its slope (m) can be calculated by using the equation (3)*

$$|leng(l_0) - leng(l_s)| < \delta \Rightarrow l_0 \text{ is straight} \quad (2)$$

$$m = (Y_{l(1)} - Y_{l(0)}) / (X_{l(1)} - X_{l(0)}) \quad (3)$$

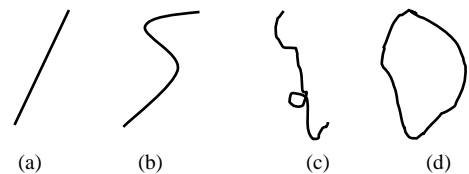


Figure 2. Collection of lines

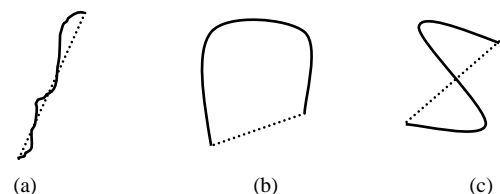


Figure 3. Different situations of the segment



Figure 4. Computing the degree of inclination

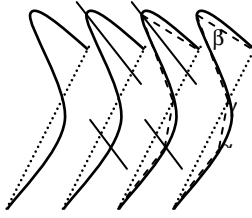


Figure 5. Line containing two detours

2) In the other case, we distinguish two states:

a) If the initial segment doesn't cross the straight line "Fig. 3" (b), then we consider that the segment contains one detour. To compute its degree of inclination we divide the segment into two segments by using the axe of the straight line. For each sub-segment we joint its endpoints and repeat the preceding steps until we reach the state when the segment can be represented by set of straight lines. "Fig. 4" shows an example of how a segment is divided. The degree of inclination is the sum of the angle (α_i) between the final straight lines. Knowing that the angle(α) between tow lines which their slopes are m_1, m_2 can be obtained by the equation (5).

$$\text{Inclination} = \sum_{i=1}^n \alpha_i \quad (4)$$

$$\alpha = \tan^{-1}\{ (m_2 - m_1) / (1 + m_2 m_1) \} \quad (5)$$

b) If the initial segment crosses the straight segment in n_0 point, then the line segment contains $n_0 + 1$ detours. For each detour we compute the degree of inclination as presented in 2.a. "Fig. 5" presented an example of a segment which contains tow detours.

We present in the "Fig. 6" our model of segment's representation by taking into account the previous different cases.

For two segments, 33 topological relationships [14] can be founded. These relations can be simplified by considering that two segments, in network, are either disjoint if they do not intersect in any point (boundaries, interiors) or meet if the segments only intersect in their boundaries. Then, two segments may only meet in one (#1-meet) or two (#2-meet) points (the segment's endpoints). If two segments intersect in another point, different from their boundaries, the two segments will be divided into new segments in a manner that the new segments have this point as an endpoint and their other endpoint is one of the initial segments' endpoints.

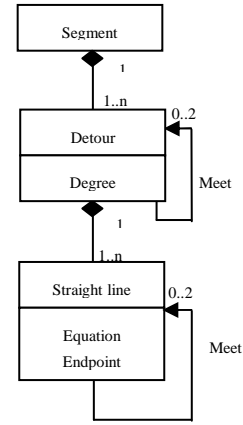
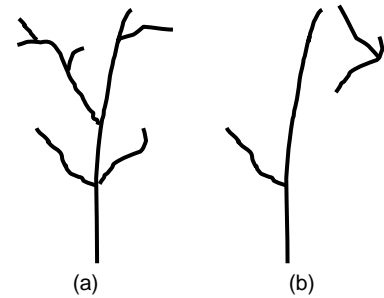


Figure 6. Model of segment representation

Definition 2: A network N is a set of connected segments S_i where all segments are exclusively related by #1-meet, #2-meet, or disjoint relations. "Fig. 7" shows examples of different configurations; the network (a) is adequate of the definition of homogeneous network whereas network (b) is not.

Each network contains a collection of segments linked by intersection point. We will use the intersection points to represent the network; knowing that each intersection point in a network must have at least three segments (if an intersection point has only two segments, these two segments will be aggregated in one segment). Around an intersection point, the connected line segments are cyclically ordered. For example, if we choose the orientation of counter-clockwise, the segments intersects in "Fig. 8" can be represented by the sequence <ABCDE>.

Two different intersection points may have the same number of segments; therefore to make difference between them we will take into account the angles among the segment. The angle between the segments can be calculated by using the equation 5. "Fig. 9" shows our model for network



representation.

Figure 7. Collection of networks

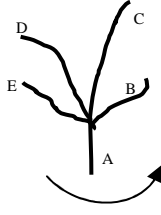


Figure 8. The orientation around intersection

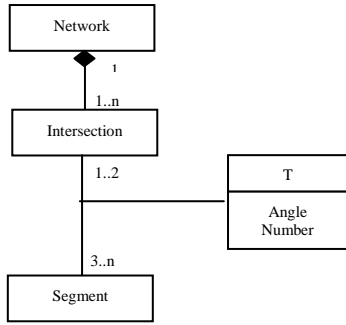


Figure 9. Model of network representation

IV. SIMILARITY OF NETWORKS

The assessment of the similarity among the networks when large changes take place is a complex task, though people do it innately. According to [16] the similarity is an intuitive and subjective judgment which displays no strict mathematical models. In this section we focus on a computational method to assess the similarity among networks.

In order to understand how people do this task, we take an example of a user who wants to retrieve the networks that resemble sketched configuration and ranks the results according to their degrees of similarity. We find that the user usually does not consider that all the information has the same importance in the network; and when he creates his sketch, he puts only the information which he considers as important and abandons the other. Therefore, it is important to provide for each component in the network a suitable weight according to its importance. Another observation is the user use approximate information not precise information. Generally quantitative information is used to describe numeric information, but it is precise information and can be changed dramatically with resolution and accuracy changes. Therefore, it is more suitable to map this information from quantitative to qualitative representation which is more robust with changes.

The network is composed of segments linked by intersection points. Identical networks have the same number of components but similar ones don't necessary have that. To assess similarity among networks we will tie a weight to each component in the network according to its importance. The importance of segment in the network generally commensurates with its length and its shape. The segment's relative length is quantitative and changes with accuracy and

resolution changes. In order to have the qualitative relative length of the segments we will use a clustering method. In this method, firstly we rank the segments' length in ascendant order. Secondly, we calculate the difference between each two following values, and the average of these differences. Thirdly, we will fix the segments' lengths which are the difference between them to see whether they are equal or bigger than the average. Finally, other segments' lengths will be clustered to fixed length provided that the interval's length is smaller than the average difference; if it is not possible to do that, we will create new fixed segments' lengths.

After tying a weight to each segment, we will tie a suitable weight to each intersection point. This weight is obtained by the sum of the segment's length which has this intersection point as an endpoint. We will use only these weights to rank the intersection points according to its importance. Thus, we don't need to map it to qualitative information. Another characteristic of intersection point is the angles among the segments which have this intersection point as endpoint. We will map it to qualitative information by dividing the full circle (4 right angles) into eight angles and will note them by $\alpha, \beta, \sigma, \tau, \upsilon, \varpi, \mu, \phi$; as it is shown in "Fig. 10":

To assess the similarity among networks we distinguish between exact matching and approximate matching. The first is adequate to the accuracy change (i.e. only the components' lengths and the angles among them have modestly changed). The second is adequate to the resolution change (i.e. some components are dropped or aggregated).

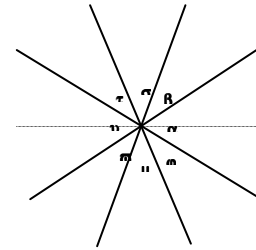


Figure 10. Dividing the full circle into eight angles

TABLEAU 1. TRANSFORMATION THE ANGLE FROM QUANTITATIVE TO QUALITATIVE INFORMATION

Quan. Angle	Qual. Angle
α	$2\pi/8 \leq \theta \leq \pi/8$
β	$\pi/8 \leq \theta \leq 3\pi/8$
σ	$3\pi/8 \leq \theta \leq 5\pi/8$
τ	$5\pi/8 \leq \theta \leq 7\pi/8$
υ	$7\pi/8 \leq \theta \leq 9\pi/8$
ϖ	$9\pi/8 \leq \theta \leq 11\pi/8$
μ	$11\pi/8 \leq \theta \leq 13\pi/8$
ϕ	$13\pi/8 \leq \theta \leq 15\pi/8$

A. Exact matching:

In this section we present a method to assess the similarity among homogeneous networks, when small changes occur. In this case the two networks have the same number of components and only the components' lengths and the angles among them have modestly changed. Therefore, we try to match the intersection points in the two networks by starting with the point which has the biggest weight in the first network with its counterpart in the second network; the two points must have the same number of segments. Secondly, we take all intersection points which have a direct link with the biggest intersection point in the first network and try to do matching with their counterpart in the second and so on. If we can't do matching between networks' intersection points according to the previous procedure we will try to do an approximate matching; otherwise we will have two trees which is used to do the matching among the segments and the angles. In this case the degree of similarity can be calculated by comparing the angles and the lengths among the asymmetric components, and it can be given by the following equation:

$$S = \omega S_L + \zeta S_A \quad (6)$$

Knowing that S_L represents the similarity between the lengths and S_A represents the similarity between the angles. If we have two networks which have n segments grouped into k clusters and m angles, S_A and S_L can be given by the following equations, knowing that Δ_i , δ_i represent respectively the difference between the lengths' class and qualitative angles between the asymmetric components.

$$S_L = 1 - \sum_{i=1}^n \frac{\Delta_i}{n * k} \text{ and } \Delta_i = \begin{cases} 0 & \text{if } |x_i - x_i'| \leq 1 \\ |x_i - x_i'| & \end{cases} \quad (7)$$

$$S_A = 1 - \sum_{i=1}^m \frac{\delta_i}{8m} \text{ and } \delta_i = \begin{cases} 0 & \text{if } |\theta_i - \theta_i'| \leq 1 \\ |\theta_i - \theta_i'| & \end{cases} \quad (8)$$

B. Approximate matching:

In this section we present a method to assess the similarity among homogeneous networks, when large changes occur. In this case the two networks don't have necessarily the same number of components. Doing the matching in this situation is more complex than exact matching because some components may disappear or aggregate. We propose an original method based on the construction of the main network.

The main network is obtained from the initial network by removing small segments which have a big possibility to drop or aggregate with accuracy and resolution changes. But by doing so, it must be ensured that the main network remains connected. To do this, firstly we keep only the segments which represent the links among intersection points. This can be done by arbitrarily taking an intersection point and keep only the segment which links this point to other intersection points. For each of these points we apply the same method and so on until we reach all intersection points. Afterward we add to this

network each segment which has a length equal or more than the average length.

After the extraction of the main network we have three cases adequate to the result. First case represents the state of the two main networks which have the same number of components. As a result we can do the exact matching over them. Second case represents the state of the two main networks which don't have the same number of components; each of them has at least one intersection point. In this case we try to do the matching among intersection points by using the angles' sequence. The problem in this case is that as some components can be dropped, the angles' sequence of intersection can be changed. With the deletion of one segment from the intersection point, the two angles formed by this segment and its neighbourhood segments will be integrated in one angle which is the sum of the two initial angles. The last case represents the state where one main network doesn't have any intersection point. In this case we try to do the matching by using the characteristic of the segment (the number of detour and the degree of inclination). The degree of similarity for the two previous situations is given by the equation (8). Knowing that K_i , L_i represent respectively the segment in the first (second) main network which has asymmetric components in the second (first) main network.

$$S = \frac{1}{2} \left(\frac{\sum_{j=1}^i K_i}{\sum_{i=1}^n S_{1,i}} + \frac{\sum_{j=1}^i L_i}{\sum_{i=1}^m S_{2,i}} \right) \quad (9)$$

Example:

Let us take the two networks in the following figure:

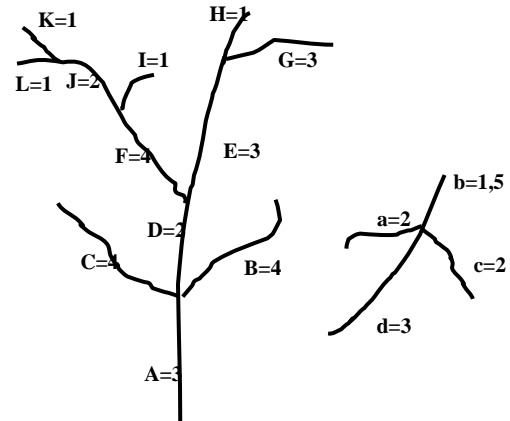


Figure 11. Comparison between two networks

We find that the two networks don't have the same number of components so we will extract their main networks.

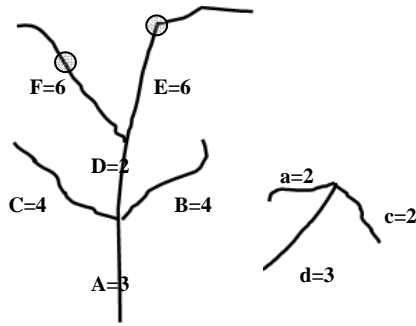


Figure 12. Main networks

The first main network in the figure (a) will be represented as following by using its intersection points: $ABDC(13, \tau, \beta, \sigma, \sigma)$, $DEF(14, \upsilon, \beta, \tau)$,

The second main network in the figure (b) will be represented as following by using its intersection point: $adc(7, \beta, \sigma, \varpi)$

As the two main networks don't have the same number of components, we try to do the matching by using the angles' sequence. Form the second main chain we have an intersection point which has three segments and the angles' sequence in it is (β, σ, ϖ) . In the first one we have tow intersection points; one of them has three segments and the other has four segments. By comparing the two intersections points which have three segments we find that the two sequences have β in common but the rest of the sequence is completely different. Therefore we try to do the matching between $ABDC$ and adc ; knowing that the first has one more segment than the second; so we look for the angle common between the two sequences. We have β , σ in the same order, the remaining angles in the second main network ϖ can be considered as the sum of the two angles σ , σ taking into account that the asymmetric of A in the second main network was dropping. The degree of similarity between the networks is given by

$$S = \frac{1}{2} \left(\frac{4 + 4 + 2}{6 + 6 + 4 + 4 + 2 + 3} + \frac{2 + 3 + 2}{2 + 3 + 2} \right) = 0,7$$

V. CONCLUSION

We presented a method to assess the similarity between homogenous networks when the large changes occur. As with large changes some network's components may disappear or aggregate, it is impossible to use only the topological equivalence. Particularly, we have extended an existing method by considering the directional and metrical properties and introducing the importance of component in assessing the similarity.

We are programming our proposal in a software which is capable of evaluating whether two pairs of networks are similar or not. In the case of similarity, it generates a report of difference. This software will be extended to work as a search engine which allows the user to draw his sketch and look for the networks which have similar structure. Future search plan

will lead to the assessment of similarity between the scenes with complex configurations including all the geographic features.

REFERENCES

- [1] J. Dettori and E. Puppo. How Generalization Interacts with Topological and Metric Structure of Maps. Seventh International Symposium on Spatial Data Handling, pages 9A.27-9A.38, 1996.
- [2] N. Tryfona and M. Egenhofer. Multi-Resolution Spatial Databases: Consistency among Networks. Sixth International Workshop on Foundations of Models and Languages for Data and Objects, pages 119-132, 1996.
- [3] M. Egenhofer and R. Franzosa. Point-Set Topological Spatial Relations. International Journal of Geographical Information Systems, pages 161-174, 1991.
- [4] M. Egenhofer and R. Franzosa. On the Equivalence of Topological Relations. International Journal of Geographical Information Systems, pages 133-152, 1995.
- [5] A. Freksa. Using Orientation Information for Qualitative Spatial Reasoning. Theories and methods of Spatio-Temporal Reasoning in Geographic Space, Springer, pages 162-178, 1992.
- [6] D. Hernández, E. Clementini, and P. Di Felice. Qualitative Distances. Third European Conference on Spatial Information Theory, pages 45-58, 1995.
- [7] B. Buttenfield. Multiple representations: Initiative 3 Specialist Meeting Report. Technical Report, National Center for Geographic Information and Analysis, UCSB Santa Barbara, 1989.
- [8] J. Carvalho-Pavia. Topological Equivalence and Similarity in Multi-Representation Geographic Databases. PhD Thesis, University of Maine, 1998.
- [9] K. Beard. How to Survive on a Single Detailed Database. Eighth AutoCarto, pages 211-220, 1988.
- [10] M. Egenhofer, E. Clementini and P. Di Felice. Evaluating Inconsistency among Multiple Representations. Sixth International Symposium on Spatial Data Handling, pages 143-160, 1994.
- [11] E. Puppo, G. Dettori. Towards a Formal Method for Multiresolution Spatial Maps. Forth International Symposium on Advances in Spatial Databases, pages 152-169, 1995.
- [12] M. Egenhofer and K. Al-Taha. Reasoning about Gradual Changes of Topological Relationships. Theory and Methods of Spatio-Temporal Reasoning in Geographic Space, pages 196-219, 1992.
- [13] M. Egenhofer and D.Mark. Modeling Conceptual Neighborhoods of Topological Line-Region Relations. Journal of Geographic Information Systems, pages 555-565, 1995.
- [14] A. Belussi, B. Catania and P. Podestà. Towards Topological Consistency and Similarity of Multiresolution Geographical Maps. Thirteenth ACM International Symposium on Advances in GIS, pages 220-229, 2005.
- [15] M. Egenhofer and J. Herring. Categorizing binary topological relationships between regions, lines and points in geographic databases. Technical report. National Center of Ggeographic Information and analysis, University of California.1990.
- [16] A. Tversky. Features of similarity. Peschological Review, pages 327-325, 1977.