A MODULARITY APPROACH TO THE DELINEATION OF FUNCTIONAL REGIONS FROM SPATIAL INTERACTION DATA

Junya FUKUMOTO, Tohoku University, fukumoto@plan.civil.tohoku.ac.jp

Yoshihiro OKAMOTO, Ibaraki Prefecture, yo.okamoto@pref.ibaraki.lg.jp

Akihito UJIIE, Tohoku University, khujiie@plan.civil.tohoku.ac.jp

ABSTRACT

Functional region is a region that is composed of areas or locational entities which have more interaction or connection with each other than with outside areas and be functionally complementary to each other (Brown and Holmes, 1971). Since 1970's many geographer and regional scientists have been discussing how to delineate functional regions. Interaction or connection among location entities can be represented as a graph, and some researchers proposed to apply graph theoretic techniques in delineating functional regions (*e.g.* Masser and Brown, 1975; Slater, 1976). In the last decade, network science made great advances boosted by the explosion of WWW and the finding of a community that is a densely connected subnetwork in the total network became one of the hottest research topics in that field. In this study, we apply a community finding method to the delineation of functional regions, and propose a new delineation method that is a combination of a modularity approach, the most popular community finding method, and a maximum-entropy spatial interaction model. The proposed method has the following characteristics: (1) the number of delineated functional regions is endogenously determined, (2) the hypothesis behind the delineation of functional regions is clear, (3) many algorithms is available, and (4) delineated functional regions are robust to the size of the study area. We apply it to the Japanese inter-municipal commuting flow data, and show its validity and enormous potential as the delineation method of functional regions.

Keywords: functional region, modularity, entropy model, commuting OD table

INTRODUCTION

Functional region (hereafter, FR) is a region that is composed of areas or locational entities which have more interaction or connection with each other than with outside areas and be functionally complementary to each other (Brown and Holmes, 1971). A typical FR is a metropolitan area. Each metropolitan area has its core district(s) and the other districts interact with the core district(s) through commuting, shopping, working, studying and so on. According to OECD (2002), 16 of 22 countries have their official FRs. In the United States, the Federal Government defined their first Urban Metropolitan Area (UMA) in 1947. A variety of statistical data has been arranged and released on those areas basis. The definition of UMA has changed several times by now. Since 2000 population census, the Core Based Statistical Area (CBSA) has been used (OMB, 2000). CBSA consists of municipalities in the core and municipalities surrounding the core. In EU, the Urban Audit database project has started, in addition to the Eurostat regional statistics database project, in 2003 (Eurostat, 2009). Travel-to-work-areas (TTWAs) in UK and Housing Market Areas (HMA) in Scotland are the other examples of the official FR.

The concept of FR is important for both academic and policy purposes. Once FRs are delineated, many kinds of statistical data are collected for each FR, and it opens the possibility of academic research or practical policy based on quantitative evidence. From the academic viewpoints, FR is important for the following purposes; (1) To understand the growth of regions and measure the speed of growth; (2) To understand the agglomeration phenomenon and measure the agglomeration of resources such as population, firm, information and so on; (3) To quantitatively compare the size and the function of different regions; (4) To determine the boundary of study area to be analyzed. From the practical policy viewpoints, FR is important for the following purposes; (1) To determine the set of local municipalities that their policies, such as employment or housing policy, must be coordinated; (2) To assess and compare the quality of life of people living in different regions; (3) To detect the underdeveloped or fragile regions. Due to these benefits of FRs, many geographers and regional scientists have been discussing how to delineate FRs since 1970's. The existing delineation methods are common in that they use the data of interaction or connection among location entities arranged in the format of OD table. Major methods can be classified into mainly two groups: algorithmic methods and rule-based methods. Algorithmic methods apply some algorithm of markov analysis, cluster analysis, factor analysis and so on (e.g. Brown and Holmes, 1971; Masser and Brown, 1975; Slater, 1976). More strongly connected local entities are merged to form a FR. The strength of interconnectedness is measured by some algorithmic criteria. Rule-based methods apply a fixed rule (or rules) to merge strongly connected local entities and to identify FRs. Most of rules are shaped reflecting our images on FRs. Therefore, the delineating procedure of the rule-based methods is more understandable, and FRs determined by these methods are unlikely to differ from our intuitional images.

Although many methods have been proposed, there is no decisive method. It is because there is no official definition of FR, and some ad hoc assumptions must be made to delineate FRs. Major drawbacks of existing methods are as follows:

1. The number of finally delineated FRs is determined by some ad hoc criteria, such as analyst's subjective judgment (for most of algorithmic methods such as cluster analysis or factor analysis) or operationally required threshold value (for most of rule-based methods);

2. The most rule-based methods are too specific to a certain problem, and cannot be applied to more than one OD data (or the results are non-comparable even if they are applied to);

3. The algorithms used in the algorithmic methods lack the theoretical foundations of modern graph theory or network science, and the possibility of extension to more advanced cases, such as overlapping FRs or hierarchical FRs, is limited;

4. Hypothesis behind delineation is not clear and there is no validation methodology, such as statistical significance test, to examine the validity of the delineated output.

In this study we pay attention to the modularity approach of community finding that has been developed in the field of network sciences during the last decade. Employing and extending the modularity approach to the delineation of FRs, we propose a new method that can overcome the above mentioned drawbacks of existing methods.

Community is defined as a densely connected subnetwork of the total network. To understand the characteristics and functions of various networks, such as brain's communication network, protein network, social network, ecological network and so on, community finding methods were required and many researchers joined in the development. Modularity approach was proposed by Girvan and Newman (2002), and immediately became the most popular and standard method of community finding. The modularity approach has the following characteristics (for details, Fortunato, 2010):

1. Community finding problem is formulated as a maximization problem, and many algorithms to solve this problem are prepared;

2. The number of finally detected communities is endogenously determined as a result of the maximization;

3. The modularity approach can be applied to many kinds of data due to the simplicity of the principle, and it is relatively easy to compare the results;

4. The hypothesis behind the community finding is clear, and some methodology to examine the validity of the output has been proposed in existing researches;

5. There are many extension to more advanced cases, such as overlapping communities, hierarchical communities and so on.

Since any OD table can be regarded as an adjacency matrix of a directed and weighted complete graph, the modularity approach is thought to be applicable to the delineation of FRs.

Comparing the drawbacks of the existing FR delineation methods and the characteristics of the modularity approach mentioned above, we can see it has a great potential as a FR delineation method.

However, as explained in Chapter 3, it is problematic to apply the standard modularity approach to the delineation of FRs. This is because the modularity approach has been developed to find communities from the general networks and does not take care of the effect of geographical distance. The spatial interactions among local entities are governed by the law of distance, and their OD tables must be regarded as the adjacency matrixes of the spatially embedded networks. To apply the modularity approach to the delineation of FRs, we must extend the standard modularity approach to incorporate the effect of geographical distance.

The objective of this study is to extend the standard modularity approach and to develop a new method of delineating FRs. We propose a new method that combines the standard modularity approach with the maximum-entropy spatial interaction model (hereafter, MESI model). Throughout the case study, we show its validity and enormous potential as a method of FRs.

The structure of this paper is as follows. In Chapter 2, the standard modularity approach is explained. In Chapter 3, the problem caused when applying the standard modularity approach to the delineation of FRs is explained. In Chapter 4, we propose a new method of the delineation of FRs that is a combination of the modularity approach of community finding and the MESI model. In Chapter 5, we apply it to the Japanese inter-municipal commuting data and show its potentials. In Chapter 6, we summarize the contribution of the paper.

MODULARITY APPROACH

Definition of the Modularity

Modularity is a measure used in the community finding literatures. Community is defined as a set of densely connected nodes in the network. A typical example of community is a clique that is developed in social network analysis and defined as a maximal complete subgraph of three or more nodes (Faust and Wasserman, 1990). In the modularity approach, a community is not defined explicitly, instead is detected by maximizing the modularity function defined as follows:

$$Q = \frac{1}{M} \sum_{c} \sum_{i,j \in c} \left[A_{ij} - P_{ij} \right], \tag{1}$$

where i and j are the suffixes of nodes, c is the suffix of a community, A_{ij} is (i, j) element of A the adjacency matrix of the observed network (*e.g.* OD matrix of commuting), M is the total sum of the elements of A, and P_{ij} is (i, j) element of P the adjacency matrix of the null network.

In the modularity approach, it is important which kind of a null network is used, because communities are detected from the comparison of the observed and the null networks. In their seminal paper, Girvan and Newman (2002) defined the null network by the configuration model of random non-directed and non-weighted graphs. In the configuration model, the nodes are connected randomly, but the degree distribution of the graph is given exogenously. Leicht and Newman (2008) extended the null network defined by the configuration model to the directed and weighted graphs. The null model is given by

$$P_{ij} = \frac{k_i^{out} k_j^{in}}{M}, \tag{2}$$

where k_i^{out} is the out-degree of node i and k_j^{in} is the in-degree of node j respectively. As we can see from equation (1) and (2), all the subset of nodes become the candidates of community and only the subsets that combinatory maximize the modularity are detected as the communities.

Algorithms of maximizing modularity

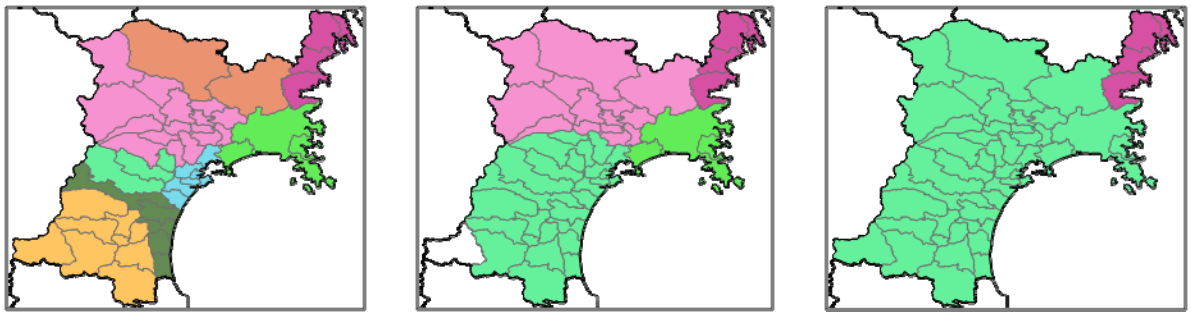
Maximization of the modularity is NP-hard problem. Many researchers have been investigating approximation algorithms and heuristics. Approximation algorithms can be classified into mainly two groups. One is an agglomerative algorithm and the other is a divisive algorithm. In the former each node is regarded as a minimal community at the beginning, then some communities are merged as the modularity increases, and this agglomerative process is iterated until the modularity cannot increase any more. The typical agglomerative algorithm is a greedy algorithm proposed by Newman (2004). In the divisive algorithm, the whole network is regarded as a maximal community at the beginning, then each community is divided into some small communities as the modularity increases and this divisive process is iterated until the modularity cannot increase any more. The typical divisive algorithm is a spectral maximization algorithm formulated in Newman (2006). As for heuristics, most of major heuristics such as taboo search, simulated annealing, genetic algorithms and so on, have been applied till now. As for the comparison of approximation algorithms and heuristics, see Lancichinetti and Fortunato (2009).

LIMITATION OF THE STANDARAD MODULARITY APPROACH

Since any OD table can be regarded as an adjacency matrix of a graph, modularity approach seems to be applicable to the delineation of FRs. However, it is problematic to apply the modularity approach with the standard null model (Equation 2) proposed by Girvan and Newman (2002). In what follows, we will ensure the problems illustrating some results of case studies and investigate the reason and action to be taken.

Case Study

Here, we apply the standard modularity approach to the Japanese inter-municipal commuting data that was surveyed in the 2005 Census. At that time, the number of total municipalities



a) Miyagi prefecture b) Tohoku region c) Whole Japan

Figure 1 Inrobustness of the standard modularity approach

was 2,264. The modularity approach was applied to the following three study areas: Miyagi prefecture, Tohoku region (including Miyagi prefecture), and the whole Japan (including Tohoku region). The numbers of local municipalities are 45, 394 and 2,264, respectively.

Figure1 shows the results of three cases. For all the cases, only the FRs in Miyagi prefecture are shown. The number of detected FRs is 8 when Miyagi prefecture is the study area, 4 when Tohoku region is the study area, and 2 when the whole Japan is the study area. We can see the detected FRs become larger as the study area expands. This result indicates that the result of modularity approach is inrobust to the choice of the size of study area.

Discussion

Girvan and Newman (2002) supposed that nodes with large degree tend to be linked with more nodes and formulated the null model based on the configuration model. The inter-municipal commuting network also has a tendency that the number of commuters from (to) a local municipality increases as the number of people living (working) there increases. The degree matters in the inter-municipal commuting network, too. But, at the same time, the effect of geographical distance cannot be ignored for the spatially embedded network. Obviously, the number of commuters has a tendency to decrease as the distance between two municipalities becomes large. There is no concern for the effect of distance in the formulation of Girvan and Newman (2002). This indicates the possibility that the inrobustness of the standard modularity approach comes from the fact of ignoring the effect of geographical distance.

We can see that the above mentioned possibility is true by investigating Equation 1 and 2. Now, suppose that study area expands from one region to two identical but far-removed regions. In this case, since two regions are far-removed and the law of distance works, the number of commuters coming in each local municipality and the number of commuters going out do not change before and after the expansion of the study area. The numbers of inter-municipal commuters within the original study area do not change, neither. However, as the study area expands, the total number of commuters is doubled. This means that, in Equation

1 and 2, only M increases and A_{ij} , k_i^{out} , k_j^{in} do not change. As a result, less dense subgraph tends to be detected as the study area expands. These results infer that it is necessary to replace the null model of the standard modularity approach with the other null model that is robust to the effect of geographical distance.

NEW METHOD

Proposal of a New Method

To overcome the limitation of the standard modularity approach and to consider the crucial effect of geographical distance, we formulate the null model as a Maximum Entropy Spatial Interaction (MESI) model. MESI model is theoretically justified by either the most-probable-state approach or the efficiency principle approach (Erlander and Stewart, 1990), and it is widely used as one of the most standard spatial interaction model. MESI model can be derived by solving the following maximization problem of entropy with the three kinds of constraints (*e.g.* generation, attraction, and total cost).

$$\max_{\{q_{ij}\}} \frac{M!}{\prod_{i,j} q_{ij}!}, \tag{3}$$

$$s.t.$$

$$\sum_i q_{ij} = k_j^{out}, \tag{4}$$

$$\sum_j q_{ij} = k_i^{out}, \tag{5}$$

$$\sum_{i,j} q_{ij} c_{ij} = C, \tag{6}$$

where q_{ij} is the spatial interaction between spatial entity i and j, and C is the total cost.

By solving the above maximization problem, the following model can be derived.

$$q_{ij} = a_i b_j k_i^{out} k_j^{in} \exp(-\gamma c_{ij}), \tag{7}$$

$$a_i = \left[\sum_j b_j k_j^{in} \exp(-\gamma c_{ij}) \right]^{-1}, \tag{8}$$

$$b_j = \left[\sum_i a_i k_i^{out} \exp(-\gamma c_{ij}) \right]^{-1}, \tag{9}$$

where a_i and b_j are the balancing factors needed to satisfy the origin and destination constraints, c_{ij} is the cost between node i and node j, γ is a parameter that means the strength of the distance decaying effect. By replacing Equation (2) with Equation (7)-(9), we formulate a new method of delineating FRs.

Our proposal is similar to the delineation method proposed by Noronha and Goodchild (1992). They pointed out the necessity of explicitly setting null hypothesis and proposed a delineation method based on the spatial interaction model. The difference between our method and theirs are the possibility of extension and the applicability of toolboxes developed in network science. It is expected that our method can be extended to more complex problems such as delineation of overlapping FRs (*e.g.* Nicosia *et al.*, 2009) or

hierarchical FRs (*e.g.* Sales-Pardo *et al.*, 2007), and can be combined with useful tools such as faster algorithms (*e.g.* Lancichinetti and Fortunato, 2009), methods of fine-tuning (*e.g.* Sun *et al.*, 2009), method of evaluating results (*e.g.* Karrer *et al.*, 2008) and so on.

In the delineation of FRs, the detected FRs are required to be connected geographically, in other words, they are required not to be separated. To incorporate the geographical connectedness condition, we add minor changes to the existing algorithms of the modularity approach. In this study, we extend the greedy algorithm.

Before formulating the algorithm, let us define some variables for the convenience. The increase of modularity when merging community i and community j is represented by $B_{ij} = A_{ij} - q_{ij}$. To incorporate the geographical connected condition, let us define B_{ij}^X as

$$B_{ij}^X = X_{ij}\left[A_{ij} - q_{ij}\right] \tag{10}$$

where X_{ij} is 1 when community i and community j are connected, 0 otherwise. Now, we can formulate the greedy algorithm for the delineation of FRs as follows;

1. All the local municipalities are regarded as a community consisting of only one local municipality. Define the set of communities by $V = \{V_1, \cdots, V_N\}$, where N is the total number of local municipalities.
2. Calculate Equation (7) for all the pairs of two communities.
3. Find the maximum pair (l, m).
4. If $B_{lm}^X + B_{ml}^X \leq 0$, then stop the algorithm.
5. If $B_{lm}^X + B_{ml}^X > 0$, then merge community V_l and community V_m.
6. Update the set of communities V, and matrix B and X.

CASE STUDY

Data

We use two kinds of data. One is the inter-municipal commuting data explained in Chapter 3. The other is the inter-municipal time distance data. This data is calculated using NITAS (National Integrated Transport Analysis System) developed by the Ministry of Land, Infrastructure, Transport and Tourism. In NITAS, for any two locational points in Japan, the minimum travelling time using car or rail can be calculated. We use a time distance between the city halls of two municipalities as the time distance of these two municipalities.

Results

We delineated two study areas. One is Miyagi prefecture and the other is Tohoku region. Figure 2 and Figure 3 shows the intermunicipal commuting flows in the two areas. The red lines represent the pairs of municipalities within that many employees commute. From Figure 2, we can see that Aoba-ku in Sendai city attracts many employees and is a center of Sendai metropolitan area. We can also see that there are some municipalities that attract some employees and work as subcentres, such as Furukawa city, Kesennuma city, Ishinomaki city and Shiraishi city. From Figure 3, we can see that about 3 to 5 centres are attracting employees in each prefectures and that inter-prefectural commuters are not many.

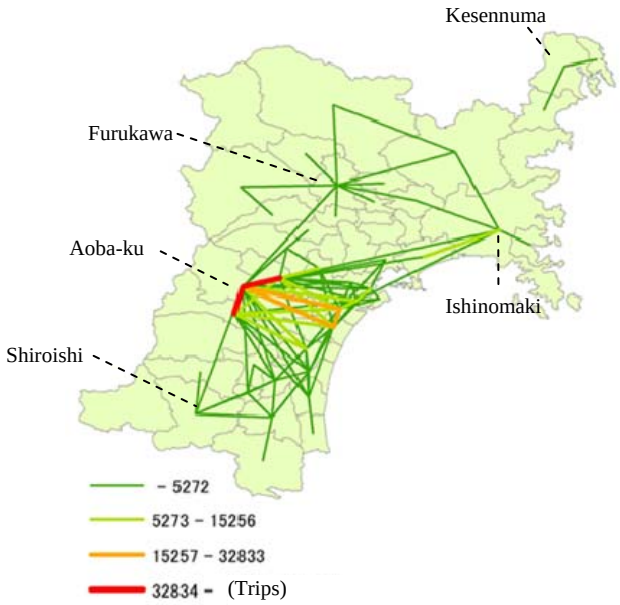


Figure 2 Commuting flow in Miyagi prefecture

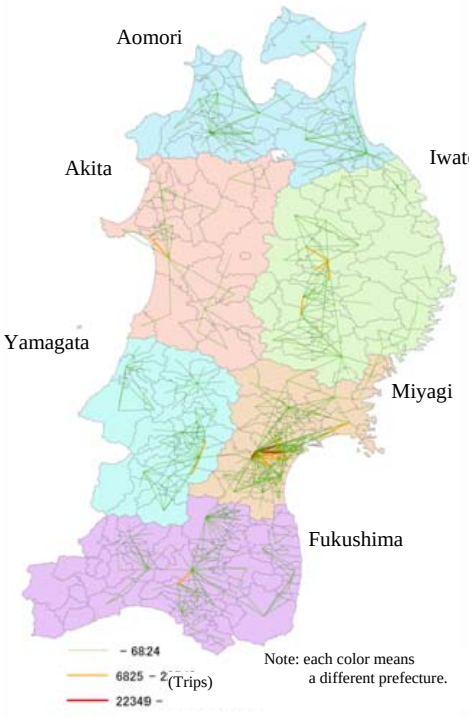
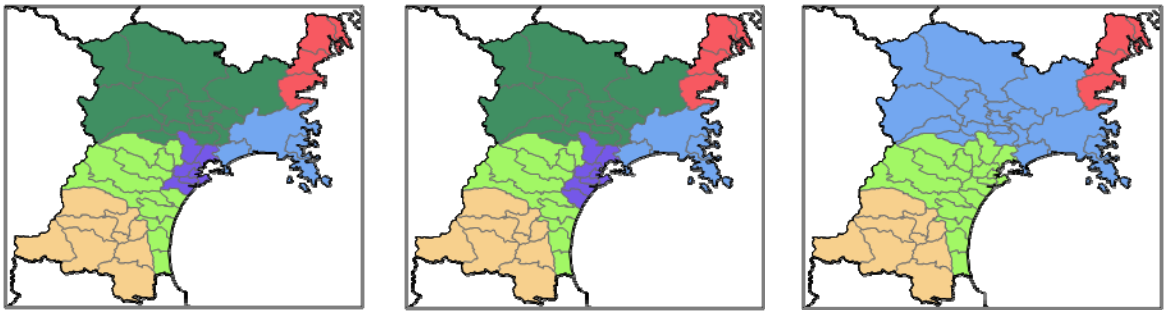


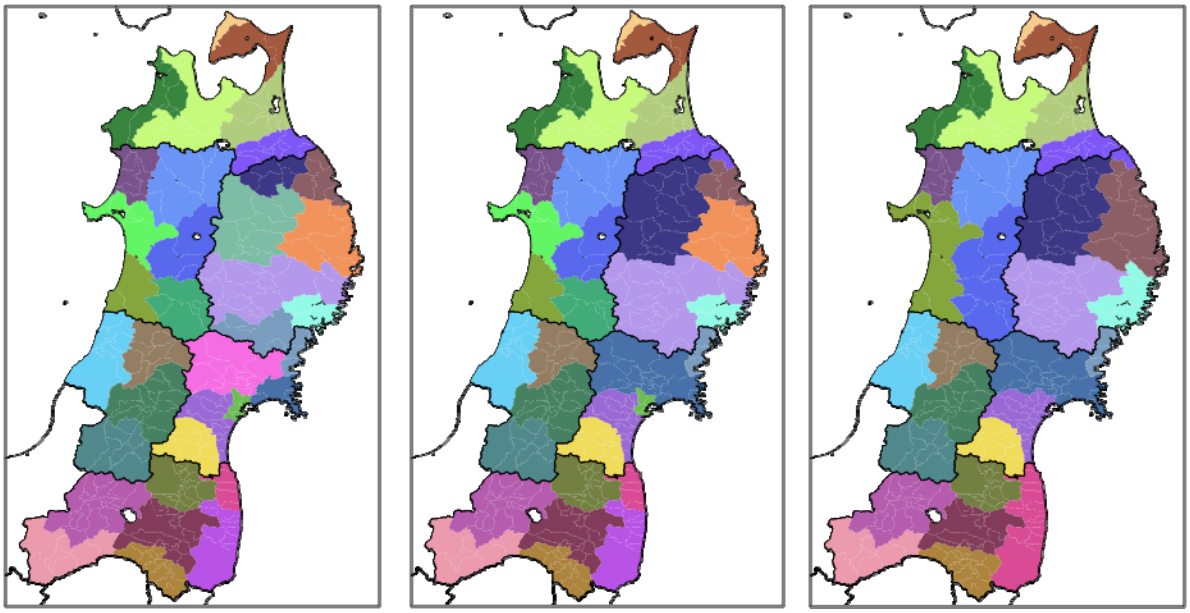
Figure 3 Commuting flow in Tohoku region

Figure 4 is the results of delineated FRs in Miyagi prefecture. For three cases shown in the figure, only the study areas differ. The distance decaying effect parameter γ is 0.04 in common. The number of detected FRs is 6 when Miyagi prefecture is the study area, 6 when Tohoku region is the study area, and 4 when the whole Japan is the study area. Though we cannot completely remove the unrobustness to the size of study area, the results become much robust compared with the result shown in Figure 1. When the study areas are Miyagi prefecture or Tohoku region, most detected FRs consists of a municipality regarded as a centre or subcentre and other municipalities surrounding it. This means that proposed method is effective in delineating FRs that are consistent with our recognition on the regional structure.



a) Miyagi prefecture b) Tohoku region c) Whole Japan

Figure 4 Robustness to the choice of study area



a) $\gamma = 0.04$ b) $\gamma = 0.035$ c) $\gamma = 0.03$

Figure 5 Unrobustness of the standard modularity approach

Figure 5 is the results of FRs in Tohoku region. For three cases shown in the figure, only the distance decaying effect parameter γ is different. The number of detected FRs is 35 when γ is 0.04, 33 when γ is 0.035, and 28 when γ is 0.03. This result shows that the proposed method is unrobust to the choice of the time decaying parameter. This parameter can be estimated by the observed OD data, and it is expected that the parameter estimates become some reference in setting the parameter. However, if we delineate FRs using the estimates of distance decaying parameter by the observed OD data, it is to be expected that the delineated FRs depend on the prediction error of null hypothetical model and that meaningless results would be derived.

To further examine the sensitivity to the choice of distance decaying parameter, we delineated Tohoku region using different values of parameter and a slightly different null

hypothesis model. We modified the null model from MESI model to a doubly constrained gravity model. Because, the latter is more general than the former, and the prediction error always become smaller by using the latter instead of the former. For reference, we estimated the distance decaying parameter by the observed OD data, and 1.52E-03 was derived.

The results of delineated FRs are shown in Figure 6. From the figure, we can see the following facts; (1) As the parameter decreases, the sizes of delineated FRs increase; (2) The result derived using the estimates from the observed OD data differ substantially from our recognition on the structure of Tohoku region; (3) The result derived using 9.2E-04 is most similar to our recognition; (4) Delineated FRs located in four of six prefectures – Aomori, Akita, Yamagata and Fukushima – are more robust to the variation of distance decaying parameter than delineated FRs located in the other prefectures (Iwate and Miyagi). As we can see from Figure 2, It is difficult to find core-periphery structure from the commuting flow pattern. The sensitivity analysis may have a potential of uncovering the degree of core-peripherial structuredness within the commuting flow pattern.

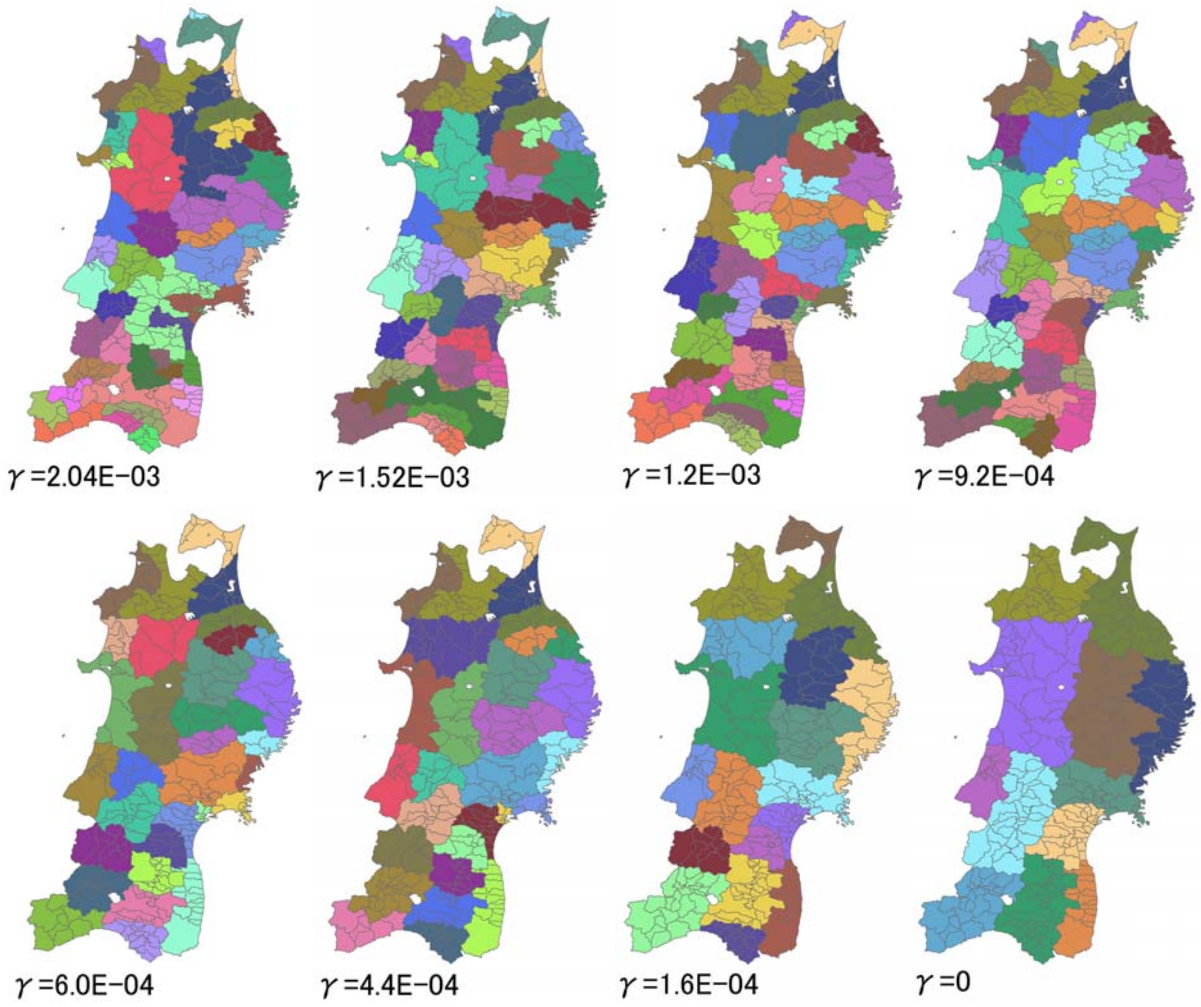


Figure 6 Sensitivity to the choice of distance decaying parameter

SUMMARY

In this paper, we proposed a new method of the delineation of FRs. The proposed method is a combination of the modularity approach and MESI model. By employing the latter model, we make it possible to apply the modularity approach to the spatially embedded network. The results of our case study show the robustness of the proposed method to the choice of study area and the sensitivity to the choice of distance decaying parameter. As we have already mentioned, the modularity approach has been extended in many directions in recent years (*e.g.* overlapping communities, communities in signed networks, those in multiplex networks and so on). This means that, by employing the updated techniques, we can extend the method proposed in this study in many directions. We want to show the result of extension in other chances.

REFERENCES

Brown, L. A. and Holmes, J. (1971), The delineation of functional regions, nodal regions, and hierarchies by functional distance approaches, J. Reg. Sci., 11, 57-72.

Claust, A., Newman, M. E. J. and Moore, C. (2004), Finding community structure in very large networks, Phys. Rev. E, 70, 066111.

Erlander, S. and Stewart, N. F. (1990) The Gravity Model in Transportation Analysis: Theory and Extensions, VSP.

Eurostat (2009), European Regional and Urban Statistics Reference Guide, 2009 Edition.

Fortunato, S. (2010) Community detection in graphs, Phys. Rep., 486, 75-174.

Girvan, M. and Newman, M. E. J. (2002), Community structure in social and biological networks, Proc. Natl. Acad. Sci. U. S. A., 99, 7821-7826.

Karrer, B., Levina, E. and Newman, M. E. J. (2008), Robustness of community structure in networks, Phys. Rev. E., 77, 046119.

Kernighan, B. W. and Lin, S. (1970), An efficient heuristic procedure for partitioning graphs, Bell Sys. Tech. J., 49, 291–307.

Lancichinetti, A. and Fortunato, S. (2009), Community detection algorithms: A comparative analysis, Phys. Rev. E., 80, 056117.

Leicht, E. A. And Newman, M. E. J. (2008), Community structure in directed networks, Phys. Rev. Lett, 1000, 118703.

Masser, I. and Brown, P. J. B. (1975), Hierarchical aggregation procedures for interaction data, Env. Plan. A, 7, 509-523.

Newman, M. E. J. (2006), Finding community structure in networks using the eigenvectors of matrices, Phys. Rev. E., 74, 036104.

Nicosia, V., Mangioni, G., Carchiolo, V. and Malgeri, M. (2009), Extending the definition of modularity to directed graphs with overlapping communities, J. Stat. Mech., P03024.

Noronha, V. and Goodchild, M. F. (1992), Modeling interregional interaction: implications for defining functional regions, Ann. Asso. Amer. Geo., 82, 86-102.

OECD (2002), Redefining Territories: Functional Regions, 042002021P1.

Office of Management and Budget (OMB) (2000), Standards of defining metropolitan and micropolitan statistical areas, Federal Register, 65, 249, December 27.

Sales-Pardo, M., Guimera, R., Moreira, A. A. and Amaral, L. A. N. (2007), Extracting the hierarchical organization of complex systems, PNAS, 104, 15224-15229.

Slater, P. B. (1976), A hierarchical regionalization of Japanese prefectures using 1972 interprefectural migration flows, Reg. Stud., 10, 123-132.

Sun, Y., Danila, B., Josic, K. And Bassler, K. E. (2009), Improved community structure detection using a modified fine-tuning strategy, Europhys. Lett., 86, 28004.

Wasserman, S. and Faust, K. (1994), Social Network Analysis: Methods and Applications, Cambridge University Press, Cambridge.