Cluster identification in maritime flows with stochastic methods

Charles Bouveyron, Pierre Latouche, Rawya Zreik, César Ducruet

Cluster identification in maritime flows with stochastic methods

Charles BOUVEYRON, Laboratoire MAP5, Université Paris Descartes, France
Pierre LATOUCHE, Laboratoire SAMM, Université Paris 1 (Panthéon-Sorbonne), France
Rawya ZREIK, Université Paris 1 (Panthéon-Sorbonne), France
César DUCRUET, CNRS & UMR 7235 EconomiX, Nanterre, France

Pre-final version of the chapter published in Ducruet C. (Ed.) (2015) Maritime Networks. Spatial *Spatial Structures and Time Dynamics*. Routledge Studies in Transport Analysis, pp. 210-228.

Since the original work of Moreno (1934), network data has become ubiquitous in computational social sciences (Snijders and Nowicki, 1997). Applications range from the study of social interactions in historical sciences (Jernite et al., 2014; Villa et al., 2008) to the analysis of maritime flows in geography (Ducruet, 2013). In particular, network analysis was applied recently to a medieval social network in Jernite et al. (2014), where the authors consider the clustering of an ecclesiastical network in Merovingian Gaul. Cluster analysis in the network context consists in grouping vertices sharing homogeneous connection profiles.

Both deterministic and probabilistic methods have been used to seek structure in these networks, depending on prior knowledge and assumptions on the form of the data. For example, Hofman and Wiggins (2008) look for specific structures called communities where nodes of the same community are preferentially connected. The alternative strategy of Handcock et al. (2007), which is a generalisation of Hoff et al. (2002), assumes the relations to be conditioned on the projection of the vertices in a social latent space. Another popular method among the community discovery approaches is based on the modularity score of Girvan and Newman (2002), though asymptotically biased (Bickel and Chen, 2009).

Most of the currently used methods derive from the stochastic block model (SBM) (Wang and Wong, 1987; Nowicki and Snijders, 2001). The SBM model assumes that each vertex belongs to a hidden cluster and that connection probabilities between pairs of vertices depend exclusively on their unobserved clusters, as in Frank and Harary (1982). In order to perform inference, standard approaches cannot be used in practice. In particular, the expectation maximisation (EM) algorithm (Dempster et al., 1977) cannot be derived because the conditional distribution of the latent groups is intractable. To overcome this issue, variational and stochastic approximations are often used. Thus, Latouche et al. (2011) used an approximation of the marginal log-likelihood, while Daudin et al. (2008) considered a Laplace approximation of the integrated classification log-likelihood. A non-parametric Bayesian approach was also proposed by Kemp et al. (2006) for estimating the number of groups while clustering the vertices.

Extensions of SBM include the mixed membership stochastic block model (MMSBM) (Airoldi et al., 2008) and the overlapping stochastic block model (OSBM) (Latouche et al., 2011). They both allow a vertex to belong to multiple clusters at the same time. More recent work focused on extending random graph models to dynamic networks (Sarkar and Moore, 2005; Xing et al., 2010; Yang et al., 2011; Dubois et al., 2013; Heaukulani and Ghahramani, 2013; Xu and Hero III, 2013), or dealing with non-binary networks such as those with weighted edges (Mariadassou et al., 2010; Soufiani and Airoldi, 2012). Some efforts were also made to take into account covariate information (Zanghi et al., 2010). For instance, the random subgraph model (RSM) (Jernite et al., 2014) was proposed to analyse directed networks with typed edges for which a partition of the vertices is available. For more details, we refer to Goldenberg et al. (2010); Salter-Townshend et al. (2012); and Matias and Robin (2014) who provided extensive reviews of statistical network models.

In this chapter, we aim to uncover clusters in a maritime flow network extracted from *Lloyd's List* where geographical information is available as well as the type of preferential commodities. The main goal of this research is to determine the possible influence of geography and cargo specialisation on the emergence of clusters in a maritime network. On the one hand, a maritime network is a multilayered system (or multigraph, multiplex graph), as different fleet types have different logics of circulation that more or less overlap and connect via port nodes. On the other, it is also a multilevel system where global, regional, and local dynamics take place simultaneously to ensure maritime freight and passenger distribution. Such an approach complements the works of Kaluza et al. (2010) and Ducruet and Zaidi (2012), which used other clustering methods and analysed the different fleets separately, without explicitly including the geographic factor in the partition. Previous works found a strong influence of geographic proximity but other possible logics remained hidden, except for hierarchical tendencies caused by hub-and-spokes configurations in container shipping. It can be hypothesised that the connections between ports are not only determined by geographic proximity but also depend on the intensity and type of circulation, but these elements need to be analysed simultaneously.

The remainder of the chapter is organised as follows. Next section presents the probabilistic model of SBM and RSM. Inference and model selection are also briefly discussed. The data and experimental setup are given afterwards, followed by an experimental comparison of both methods on the Lloyd's dataset, highlighting the advantages of using such techniques.

Probabilistic models for network clustering

This section presents the stochastic block model and the random subgraph model. Inference and model selection are also briefly discussed.

Context and notations

We consider a directed graph G with N vertices represented by its N × N adjacency matrix X. Each edge X_{ij}, describing the relation between the vertices i and j, takes its values in a finite set {0, . . . , C}. Note that $X_{ij} = 0$ corresponds to the absence of an edge. We assume that G does not have any self-loop, and therefore the entries X_{ii} will not be taken into account. In RSM, a partition P of the vertices into S classes is also assumed to be available. In both cases, our goal is to cluster the network into K groups with homogeneous connection profiles, i.e. estimating a binary matrix Z such that Zik = 1 if vertex i belongs to cluster k, 0 otherwise.

[Figure 12.1 here]

The stochastic block model

The SBM model associates to each vertex of a network a latent variable Z_i drawn from a multinomial distribution:

$$\mathbf{Z}_i \sim \mathcal{M}(1, \boldsymbol{\alpha}),$$

where α denotes the vector of class proportions. As in other standard mixture models, the vector Z_i sees all its components set to zero except one such that $Z_{ik} = 1$ if vertex i belongs to class q. The model then verifies:

$$\sum_{k=1}^{K} Z_{ik} = 1, \ \forall i \in \{1, \ldots, N\}, \tag{1}$$

and

$$\sum_{k=1}^{K} \alpha_k = 1, \tag{2}$$

where K denotes the number of components (clusters) of the mixture. Finally, the edges of the network are drawn from a Bernoulli distribution:

$$X_{ij} | \{Z_{ik} Z_{jk} = 1\} \sim \mathcal{B}(\pi_{kl}),$$

where Π = (πkl)kl is a K × K matrix of connection probabilities. According to this model, the latent variables Z_1, . . . , ZN are iid and given this latent structure, all the edges are supposed to be independent. Note that SBM was originally described in a

more general setting, allowing any discrete relational data. However, in the following we concentrate on binary edges only, i.e. C = 1.

[Figure 12.2 here]
[Table 12.1 here]

Figure 12.1 presents an example of an SBM network made of nine nodes split into three groups (indicated by the colours). As one can see, on this specific example, the within-cluster connexion probabilities (π••, π••, and π••) seem to be rather large since the nodes of each group are well interconnected. Conversely, connections between nodes of different groups are less frequent, which is due to low values of the between-group connexion probabilities (π••, π••, π••, ...). Such a matrix Π corresponds to networks made of communities, which are frequent in social networks, for instance. However, situations where between-group connexion probabilities are larger than within-group connexion probabilities are possible with the SBM model. This type of network is, for instance, very frequent in biology when studying networks of genes. Indeed, it is of interest in such a context to characterise, on the one hand, the regulatory genes and, on the other, the regulated genes. The left panel of Figure 12.2 presents the graphical model associated with SBM, and Table 12.1 summarises the model notations.

The random subgraph model

We consider now the directed graph G with a known partition P of the vertices into S classes, where each edge X_{ij} is categorical, i.e. takes its values in a finite set {0, . . . , C}. Contrary to SBM, RSM can deal with networks where C > 1. In order to simplify the notations when describing the model, we also consider the binary matrix A with entries A_{ij} such that $A_{i,j} = 1 \quad \square \Rightarrow \quad X_{i,j} ;= 0$. We also emphasize that the observed partition P induces a decomposition of the graph into subgraphs where each class of vertices corresponds to a specific subgraph. We introduce the variable s_i which takes its values in {1, . . . , S} and is used to indicate to which of the subgraphs vertex i belongs, for $i \in \{1, \ldots, N\}$.

The data is assumed to be generated in three steps. First, the presence of an edge from vertex i to vertex j is supposed to follow a Bernoulli distribution whose parameter depends on the subgraphs s_i and s_j only:

$$A_{i,j} \sim \mathcal{B}(\gamma_{s_i,s_j}).$$

Each vertex i is then associated to a latent cluster with a probability depending on si. In practice, the variable Z_i is drawn from a multinomial distribution:

$$\mathbf{Z}_i \sim \mathcal{M}(1; \boldsymbol{\alpha}_{s_i}),$$

Where

$$\forall s \in 1, \ldots, S, \sum_{k=1}^{K} \alpha_{sk} = 1.$$

A notable point of the model is that we allow each subgraph to have different mixing proportions αs for the latent clusters. We denote hereafter α = (α1, . . . , αS). Finally, if an edge between i and j is present, i.e. $A_{ij} = 1$, its type X_{ij} is sampled from a multinomial distribution with parameters depending on the latent clusters. Thus, if i belongs to cluster k and j to cluster l:

$$X_{i,j} | Z_{ik} Z_{jl} = 1, A_{ij} = 1 \sim \mathcal{M}(1, \boldsymbol{\pi}_{kl}),$$

where the sum over the C types of each vector πkl = (πkl1, . . . , πklC) is:

$$\forall (k, l) \in \{1, \ldots, K\}^2, \sum_{c=1}^{C} \pi_{klc} = 1.$$

If there is no edge between the two vertices, the entry X_{ij} is simply set to $X_{ij} = A_{ij} = 0$.

We point out that the choice of separating the role of the known subgraphs and the latent clusters was originally motivated by a parsimony concern. An alternative approach would consist in allowing the presence of an edge and its type to depend on both the subgraphs and latent clusters. However, this would dramatically increase the number of model parameters to be estimated. Indeed, for a network with S = 6, K = 6, and C = 4, it would require K 2S2(C + 1) + SK = 6516 parameters while RSM only involves S2 + K 2C + SK = 216 parameters.

The right panel of Figure 12.2 presents the graphical model associated with RSM, and Table 12.1 summarizes the model notations. Figure 12.2 allows us to see the conceptual differences between the SBM and RSM models. In particular, the specific role of A_{ij} appears here clearly.

[Figure 12.3 here]

Figure 12.3 presents an example of an RSM network made of nine nodes belonging to two subgraphs (denoted through the form of nodes) and split into three groups (indicated by the colours). A main difference with an SBM network is, of course, the presence of several (here C = 2) types of edges and a partition, assumed to be known,

of the network into (here S = 2) subgraphs. In this specific case of a network generated according to the RSM model, the probability of an edge relies on two different parameters: γ and Π. The parameter γ governs the possibility of an edge (of any type) between two nodes, and this depends on the subgraphs they belong to. For instance, the presence of an edge between nodes 2 and 7 relies on γQD. Then, if the method decides that an edge exists between those two nodes, the type of the edge is drawn from a multinomial distribution with probabilities π••. Let us recall that π•• is here a vector of C probabilities. It turns out that the specific edge between nodes 2 and 7 was chosen to be of the type "continuous line" and not of the type "dashed line".

Inference and model selection

Given a network, the inference task consists in looking for estimates of the model parameters and cluster memberships. SBM and RSM fall in the family of mixture models for which the expectation maximization (EM) algorithm is the standard inference procedure (Dempster et al., 1977). It is an algorithmic procedure which iteratively maximizes the likelihood relying on the expected complete data likelihood (McLachlan and Krishnan, 1997). Unfortunately, the EM algorithm depends on the conditional distribution of the cluster membership matrix Z given the network, which is here intractable. As an alternative, variational approaches can be used to derive an approximate inference scheme (see for instance Jordan et al., 1999). The key point is to approximate the conditional distribution of Z by assuming the conditional independence of Z1, . . . , ZN . The corresponding algorithm is called variational EM (VEM) (Daudin et al., 2008). Note that an alternative strategy consists in focusing on the optimization of the complete data likelihood (Zanghi et al., 2008). This strategy is often called classification EM (CEM). In this case, the choice of the number K of latent groups cannot be based on the observed likelihood, which is not tractable, but can be done using criteria such as the integrated classification likelihood (ICL) criterion (Daudin et al., 2008).

In order to perform model selection, it is also possible to consider the two models presented above in a Bayesian framework. The principle is to see the model parameters as random variables and to make assumptions on their distributions. In practice, conjugate prior distributions are chosen to simplify the inference, which can be done either by relying on sampling techniques (as used in Nowicki and Snijders, 2001), such as Markov chain Monte Carlo (MCMC), or Bayesian extensions of the VEM algorithm (Latouche et al., 2012; Jernite et al., 2014). The latter, called variational Bayes EM (VBEM), is preferred in the context of networks for scaling reasons. Contrary to VEM which maximizes an approximation of the likelihood, VBEM focuses on an approximation of the marginal likelihood where all model parameters and cluster memberships are integrated out. Alternative strategies rely on allocation sampler (Mc Daid et al., 2013) or greedy search (Côme and Latouche, 2015).

The model selection, which mainly consists here in choosing the appropriate number K of groups, can be done afterwards by considering the approximate marginal likelihood. Thus, K is chosen such that the corresponding criterion is maximized.

Application to the maritime network

This section now focuses on the application of both the SBM and RSM methods to a maritime flow network, extracted from the well-known *Lloyd's List*.

The Lloyds' *data*

Data was obtained from the printed *Lloyd's Voyage Record* published in October and November 2004, which details for each merchant vessel its successive movements from one port to another. Four main types of vessels are retained (containers, solid bulk, liquid bulk, and passengers/vehicles) calling at the world's 500 largest ports based on their degree (number of connected neighbours in the graph), from the original 2,737 ports or 1,815 port cities referenced in the original dataset, and are complemented by additional sources to retrieve their tonnage capacity (see Ducruet, 2013). Each port was assigned to a large region or continent, namely Asia, Europe/Mediterranean, North America, Latin America, Oceania/Pacific, Middle East/Red Sea, and Africa.

Experimental setup

From the raw database of vessel flows, we constructed an adjacency matrix between ports as follows: first, for every pair of ports, we considered the total tonnage of each commodity type by summing overall ship movements between those ports; second, we retained the main commodity type associated to each pair and drew an edge of the corresponding type between the two ports. In practice, the adjacency matrix contains entries ranging from 0 (no movements) to 4 (for the 4 commodity types taken into account). Figure 12.4 presents the adjacency matrix where commodity types are denoted using colours.

[Figure 12.4 here]

In order to apply the SBM and RSM model to those data, we used the mixer and Rambo packages for the R software. The mixer package (version 1.8) implements the VEM, VBEM, and CEM algorithms for SBM. We used the latter method for the inference, mainly because of its scaling properties. The package also allows the selection of an appropriate number K of groups for the data at hand with two criteria, and provides insightful visualisations of the results. We considered the ICL criterion, which is the one available for the CEM algorithm. On the other hand, the Rambo package (version 1.1) proposes the VBEM algorithm for the inference of the RSM model. Model selection is considered based on an approximation of the

marginal likelihood. Some meaningful plots allow the visualisation of the clustering results as well. Both algorithms were run for a number K of clusters ranging from 2 to 20 and, for each method, the value of K was chosen such that the associated criterion was maximum.

Analysis of the network with SBM

We first analyse the network with SBM. Listing 1 gives the command lines in the R language to run mixer on a binary version of the adjacency matrix (named hereafter X).

Listing 1: Analysis of the network with SBM

```
# Loading the librarylibrary(mixer)
# Binarization ofX
X = as.numeric(X != 0)
# Clustering with mixer
res.sbm = mixer(X,qmin=2,qmax=20,method="classification")
# Visualization ofthe clustering resultsplot(res.sbm)
# Selection ofthe bestSBM modelresBest= getModel(res.sbm)
```

Figure 12.5 presents the output of the mixer function. As one can observe, the ICL criterion peaks at K = 10, meaning that an appropriate number of groups for this network seems to be 10 groups for the SBM model. The reorganised adjacency matrix allows us to see different kinds of groups. First, the network comprises one large and sparse group (cluster 1) of ports with few connections. Second, the nine other groups have much larger intra-connection probabilities. We can also note the presence of clusters which tend to connect with nodes of other clusters. Those ports can be considered as hubs. As shown in Figure 12.6, except for cluster 1, most clusters correspond to geographical regions. For instance, it appears that clusters 2 and 3 can be associated with the Europe/Mediterranean and/or North American regions while clusters 4 and 6 mainly include ports from the Oceania/Pacific and Asia regions. Some other clusters, such as cluster 9, are made of hubs which allow ports of different geographic regions to connect.

[Figure 12.5 here]

Interestingly, the original SBM model cannot take a priori geographical information into account. Moreover, it only focuses on binary edges and cannot deal with categorical relationships. Nevertheless, by only looking at the presence and absence of flows between ports, one can see that the geographic information is retrieved. This tends to show that the organisation of the maritime network is mainly explained by the geography where the domination and competition between ports occurs within regions.

Analysis of the network with RSM

We then used the Rambo package to cluster the data according to the RSM model. In this case, the adjacency matrix with categorical entries was used in addition to the partition of the ports by continents. Since we provide geographic information about the network nodes to the algorithm, it should be able to focus on other patterns hidden in the data. Listing 2 gives the command lines in the R language to run the RSM function on the adjacency matrix.

Figure 12.7 presents the output of the RSM function. The variational approach finds K = 5 clusters. The reorganised adjacency matrix according to the clusters found is given in Figure 12.8. We observe that clusters can be associated with the use of specific types of vessels. Thus, cluster 2 is made of ports which tend to connect mainly through passenger/vehicle vessels. Moreover, cluster 3 contains ports interacting through solid bulk vessels. Similarly, cluster 4 can be associated with containers and cluster 5 with liquid bulk vessels. Interestingly, there are no strong connection profiles from ports in different clusters. We could have expected ports interacting through a type of vessels to have notable connections with ports associated to other types of vessels. Overall, this is not the case. The clusters found are essentially defined by their ports interacting through specific types of vessels.

[Figure 12.6 here]
[Figure 12.7 here]

As shown in Figure 12.9, the clusters found are not related to geographic regions, although cluster 3 mainly contains ports from the Europe/Mediterranean region. This is the key advantage of relying on RSM rather than SBM. Since the geographic information is given a priori, the clustering technique for RSM can uncover other patterns present in the data. Here, the results highlight that all the regions are organised through clusters of interacting types of vessels. Moreover, we point out that this methodology allows, by removing the geographical factors present in the data (as shown in the previous section), to assign roles to ports depending on maritime flows.

Listing 2: Analysis of the network with RSM

```
# Loading the librarylibrary(Rambo)
# Clustering with mixer
res.rsm  = rsm(Z,sub,Klist=2:6,nbredo=1,maxit=50,disp=TRUE)
# Visualization ofthe clustering resultsplot(res.rsm)
# Selection ofthe bestSBM  modelresBest= res.rsm$output[[5]]
```

[Figure 12.8 here]
[Figure 12.9 here]

Conclusion

In this chapter, we considered the SBM and RSM models for the clustering of ports in a maritime network created from the printed *Lloyd's Voyage Record* published in October and November 2004. We pointed out the advantages and the flexibility of the two models, and gave a short review of some of the existing approaches for their inference. In particular, we mentioned the use of variational approximations to derive tractable quantities. The two models gave rise to different and complementary results. By only looking at the presence or absence of connections between ports (SBM), we retrieved mainly geographical clusters highlighting the influence of physical geography and regional markets on port competition and maritime network configuration. However, using the known regions as subgraphs (RSM) and taking into account fleet types uncovered other types of clusters hidden in the data. These clusters are made of ports interacting through specific types of vessels. This allows us to assign a role to each port, by removing the geographical factors.

Acknowledgements

The authors would like to thank Olivier Joly, Marine Le Cam, and Brahim Ould Ismail at UMR IDEES (Le Havre) for their help on data provision and preparation.

References

Airoldi E.M., Blei D.M., Fienberg S.E., Xing, E.P. (2008) Mixed membership stochastic blockmodels. *The Journal of Machine Learning Research*, 9**:** 1981-2014.

Bickel P.J., Chen A. (2009) A nonparametric view of network models and Newman–Girvan and other modularities. *Proceedings of the National Academy of Sciences*, 106(50): 21068-21073.

Côme E., Latouche, P. (2015) Model selection and clustering in stochastic block models with the exact integrated complete data likelihood. *Statistical Modelling* (in press).

Daudin J.J., Picard F., Robin S. (2008) A mixture model for random graphs. *Statistics and Computing*, 18(2): 173–183.

Dempster A.P., Laird N.M., Rubin D.B. (1977) Maximum likelihood for incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, B39: 1-38.

Dubois C., Butts C.T., Smyth P. (2013) Stochastic blockmodelling of relational event dynamics. Proceedings of the *International Conference on Artificial Intelligence and Statistics*, Journal of Machine Learning Research Proceedings, 31: 238-246.

Ducruet C. (2013) Network diversity and maritime flows. *Journal of Transport Geography*, 30: 77-88.

Ducruet, C., Zaidi, F. (2012) Maritime constellations: A complex network approach to shipping and ports. *Maritime Policy and Management* 39(2): 151-168.

Frank O., Harary F. (1982) Cluster inference by using transitivity indices in empirical graphs. *Journal of the American Statistical Association*, 835-840.

Girvan M., Newman M.E.J. (2002) Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12): 7821-7826.

Goldenberg A., Zheng A.X., Fienberg S.E. (2010) *A Survey of Statistical Network Models*. Now Publishers.

Handcock M.S., Raftery A.E., Tantrum, J.M. (2007) Model-based clustering for social networks. *Journal of the Royal Statistical Society: Series A,* 170(2): 301-354.

Heaukulani C., Ghahramani Z. (2013) Dynamic probabilistic models for latent feature propagation in social networks. *Proceedings of the 30th International Conference on Machine Learning (ICML-13), pp. 275-283.*

Hoff P.D., Raftery A.E., Handcock M.S. (2002) Latent space approaches to social network analysis. *Journal of the American Statistical Association*, 97(460): 1090-1098.

Hofman J.M., Wiggins C.H. (2008) Bayesian approach to network modularity. *Physical review letters*, 100(25): 258701.

Jernite Y., Latouche P., Bouveyron C., Rivera P., Jegou L., Lamasse S. (2014) The random subgraph model for the analysis of an ecclesiastical network in merovingian Gaul. *The Annals of Applied Statistics*, 8(1): 377-405.

Jordan M.I., Ghahramani Z., Jaakkola T.S., Saul L.K. (1999) An introduction to variational methods for graphical models. *Machine learning*, 37(2): 183-233.

Kaluza P., Kölzsch A., Gastner M.T., Blasius, B. (2010) The complex network of global cargo ship movements. *Journal of the Royal Society Interface*, 7(48): 1093-1103.

Kemp C., Tenenbaum J.B., Griffiths T.L., Yamada T., Ueda N. (2006) Learning systems of concepts with an infinite relational model. *Proceedings of the National Conference on Artificial Intelligence*, 21: 381.

Latouche P., Birmele E., Ambroise C. (2011) Overlapping stochastic block models with application to the French political blogosphere. *Annals of Applied Statistics*, 5(1): 309-336.

Latouche P., Birmele E., Ambroise C. (2012) Variational Bayesian inference and complexity control for stochastic block models. *Statistical Modelling*, 12(1): 93-115.

Mariadassou M., Robin S., Vacher C. (2010) Uncovering latent structure in valued graphs: A variational approach. *Annals of Applied Statistics*, 4(2) 715-742.

Matias C. Robin S. (2014) Modeling heterogeneity in random graphs through latent space models: A selective review. *Esaim Proc. and Surveys, 47 :55-74*.

Mc Daid A., Murphy T.B., Frieln N., Hurley N.J. (2013) Improved Bayesian inference for the stochastic block model with application to large networks. *Computational Statistics and Data Analysis*, 60**:** 12-31.

McLachlan G., Krishnan T. (1997) *The EM Algorithm and Extensions*. New York: John Wiley.

Moreno J.L. (1934) *Who Shall Survive? A New Approach to the Problem of Human Interrelations*. Nervous and Mental Disease Publishing Co.

Nowicki K., Snijders T.A.B. (2001) Estimation and prediction for stochastic blockstructures. *Journal of the American Statistical Association*, 96(455): 1077–1087.

Salter-Townshend M., White A., Gollini I., Murphy T.B. (2012) Review of statistical network analysis: Models, algorithms, and software. *Statistical Analysis and Data Mining*, 5(4): 243-264.

Sarkar P., Moore AA.W. (2005) Dynamic social network analysis using latent space models. *ACM SIGKDD Explorations Newsletter*, **7**(2): 31-40.

Snijders T.A.B., Nowicki K. (1997) Estimation and prediction for stochastic blockmodels for graphs with latent block structure. *Journal of Classification*, 14(1): 75-100.

Soufiani H.A., Airoldi E.M. (2012) Graphlet decomposition of a weighted network. *Arxiv preprint*

arXiv:1203.2821.
Villa N., Rossi F., Truong Q.D. (2008) Mining a medieval social network by kernel SOM and related methods. Arxiv preprint arXiv:0805.1374.
Wang Y.J., Wong G.Y. (1987) Stochastic blockmodels for directed graphs. *Journal of the American Statistical Association*, 82: 8-19.
Xing E.P., Fu W., Song L. (2010) A state-space mixed membership blockmodel for dynamic network tomography. *The Annals of Applied Statistics*, 4(2): 535-566.
Xu K.S., Hero III A.O. (2013) Dynamic stochastic blockmodels: Statistical models for time-evolving networks. *Proceedings of the 6th International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction*, pp. 201-210.
Yang T., Chi Y., Zhu S., Gong Y., Jin R. (2011) Detecting communities and their evolutions in dynamic social networks: a Bayesian approach. *Machine Learning*, 82(2): 157-189.
Zanghi H., Ambroise C., Miele V. (2008) Fast online graph clustering via Erdos-Renyi mixture. *Pattern Recognition*, 41: 3592-3599.
Zanghi H., Volant S., Ambroise C. (2010) Clustering based on random graph model embedding vertex features. *Pattern Recognition Letters*, 31(9): 830-836.

	Notations	Description
SBM & RSM	$\mathbf{X}$	Adjacency matrix. $X_{ij} \in \{0, 1\}$ (SBM) or $\{0, \ldots, C\}$ (RSM)
	$\mathbf{Z}$	Binary matrix. $Z_{ik} = 1$ indicates that i belongs to cluster k
	N	Number of vertices in the network
	K	Number of latent clusters
	$\boldsymbol{\alpha}$	α_k is the prior probability of cluster k
	Π	π_{kl} is the connexion probability between cluster k and l
RSM	$\mathbf{A}$	Binary matrix. $A_{ij} = 1$ indicates the presence of an edge
	S	Number of subgraphs
	C	Number of edge types
	$\boldsymbol{\gamma}$	γ_{rs} probability of having an edge between vertices of subgraphs r and s

Table 12.1: Summary of the notations used

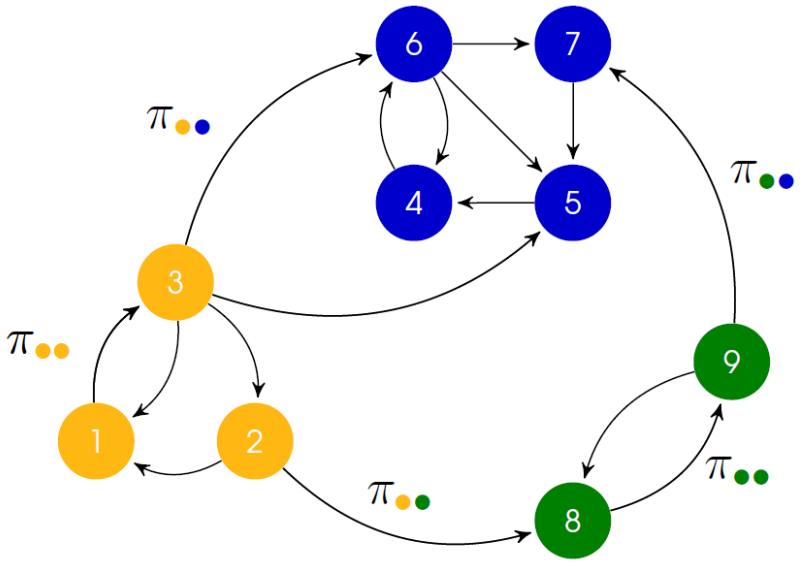

Figure 12.1: An example of a SBM network

N.B. The network is made of 9 nodes split into 3 groups (indicated by the colors). According to the SBM model, the directed edges between the nodes are assumed to be drawn from a Bernoulli distribution with probability πkl , where k, l are here colors

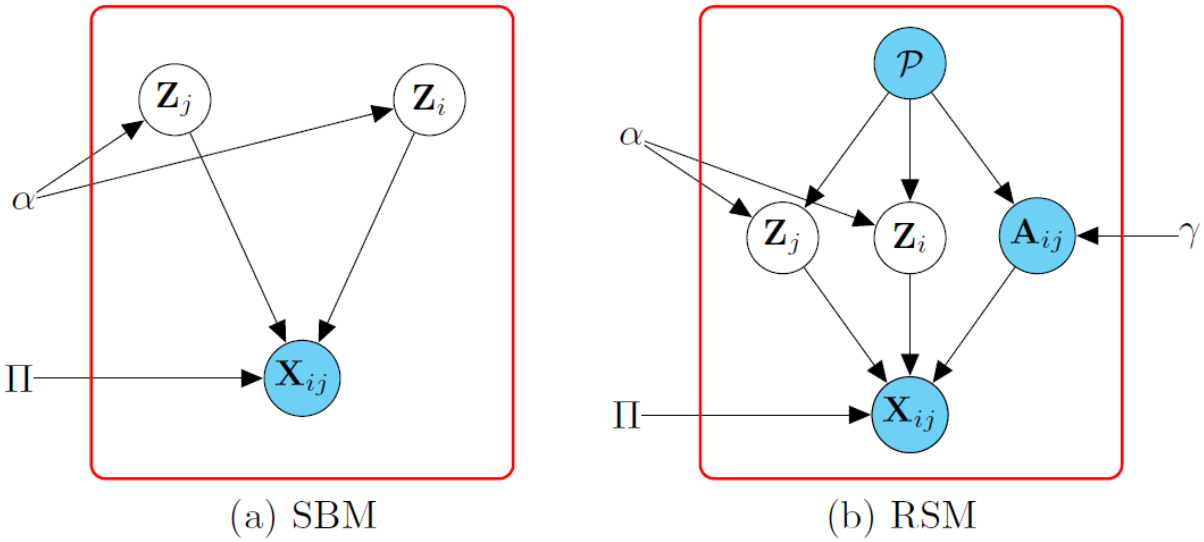

Figure 12.2: Graphical models of SBM and RSM (in their frequentist form)

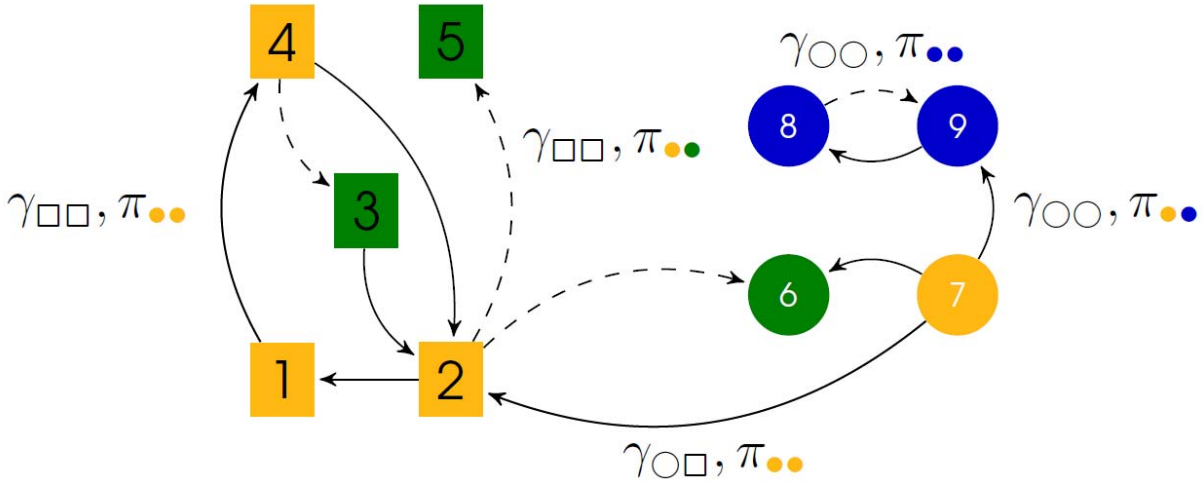

Figure 12.3: An example of an RSM network

N.B. The network is made of 9 nodes belonging to 2 subgraphs (denoted through the form of the nodes) and split into 3 groups (indicated by colors). According to the RSM model, the directed edges between the nodes can be of different types (2 types are considered here). The presence of an edge depends on the connexion probabilities between subgraphs (γ) and on connexion probabilities between groups (Π).

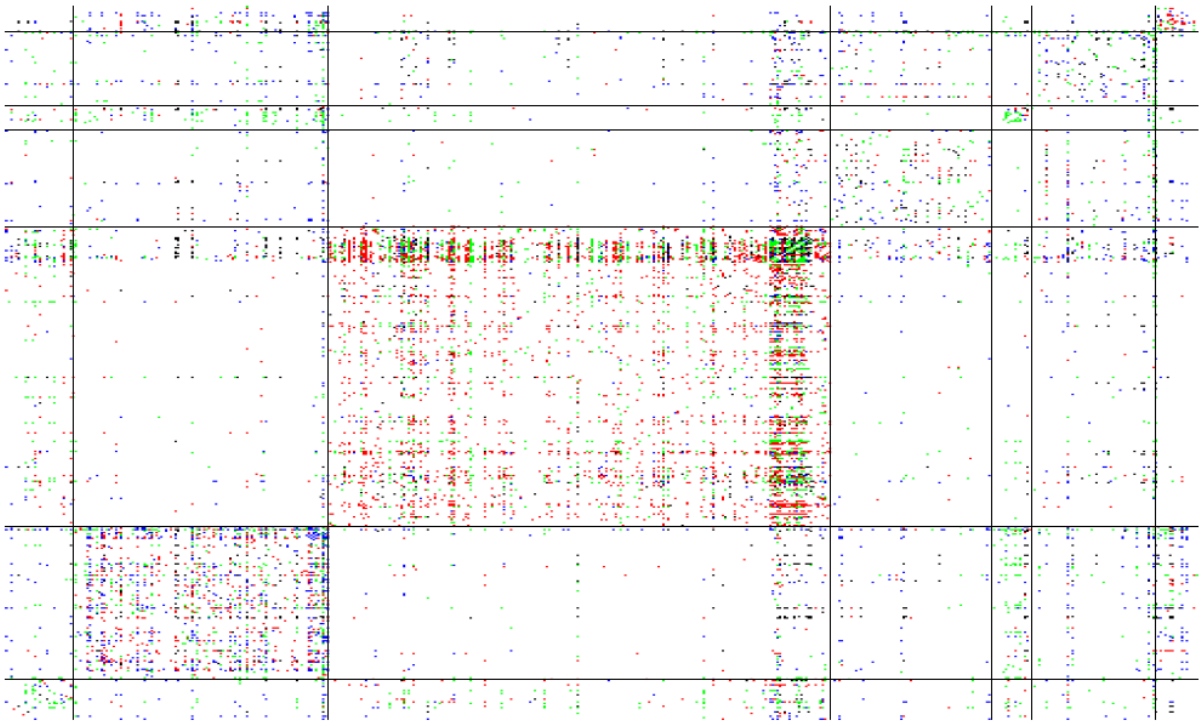

Figure 12.4: Adjacency matrix of the Lloyds data organised by continent with categorical edges: containers (black), solid bulk (red), liquid bulk (green), and passengers (blue).

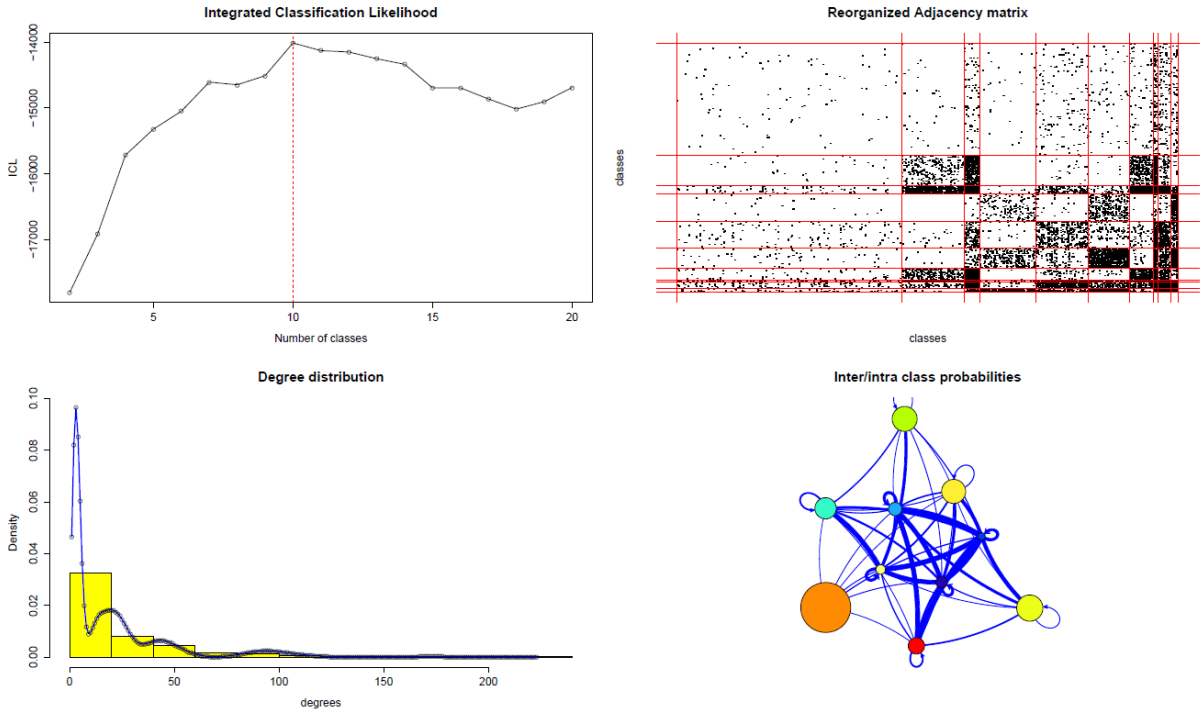

Figure 12.5: Outputs of the mixer function: values of the ICL criterion for the different values of K (top left), reorganised adjacency matrix according to the partition into 10 groups found by mixer (top right), empirical (yellow) and estimated (blue) distributions of the node degrees (bottom left), and network summarizing the relationships between the found latent clusters (bottom right)

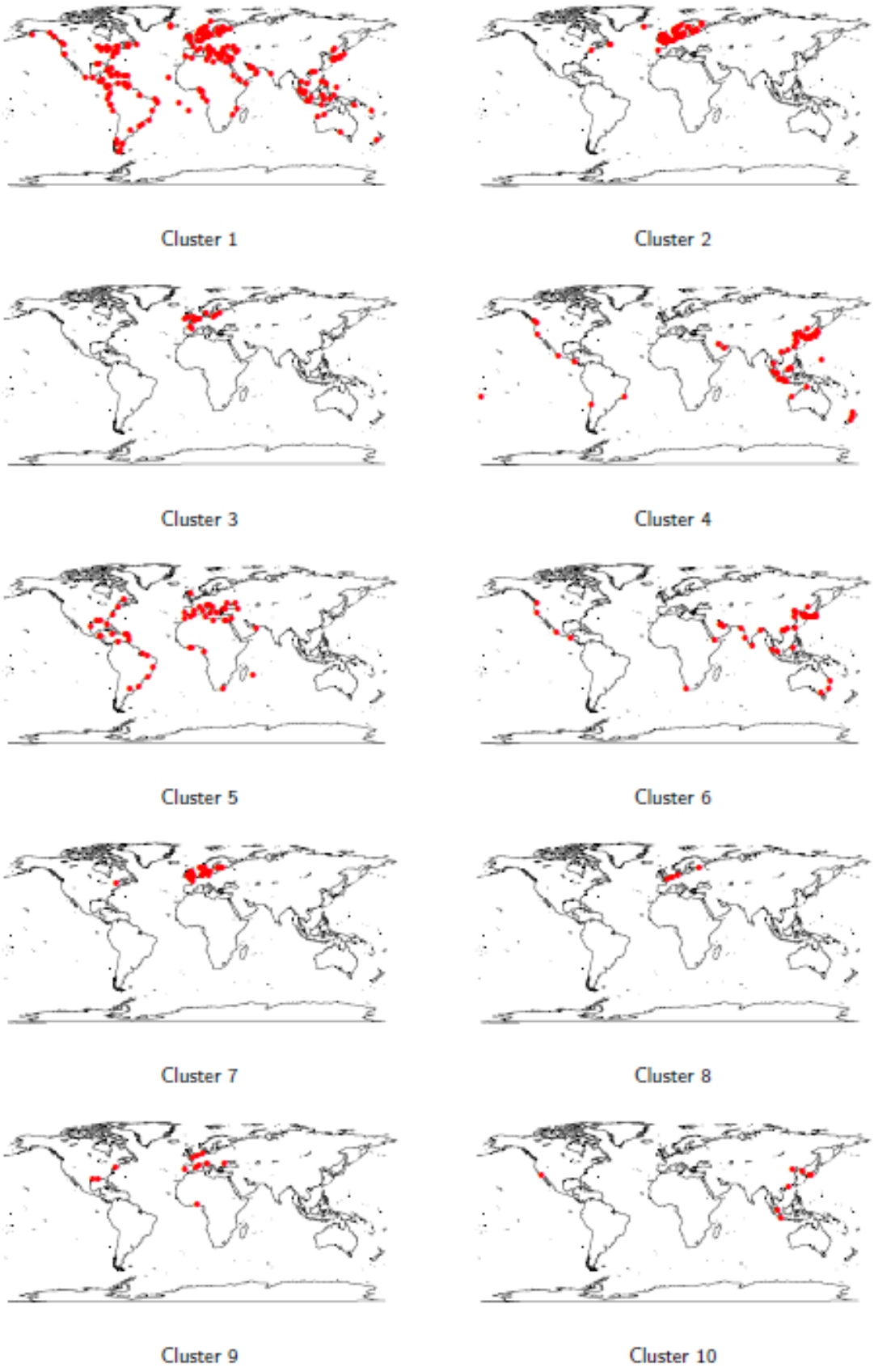

Figure 12.6: Geographic distribution of the 10 clusters found by mixer for the SBM model

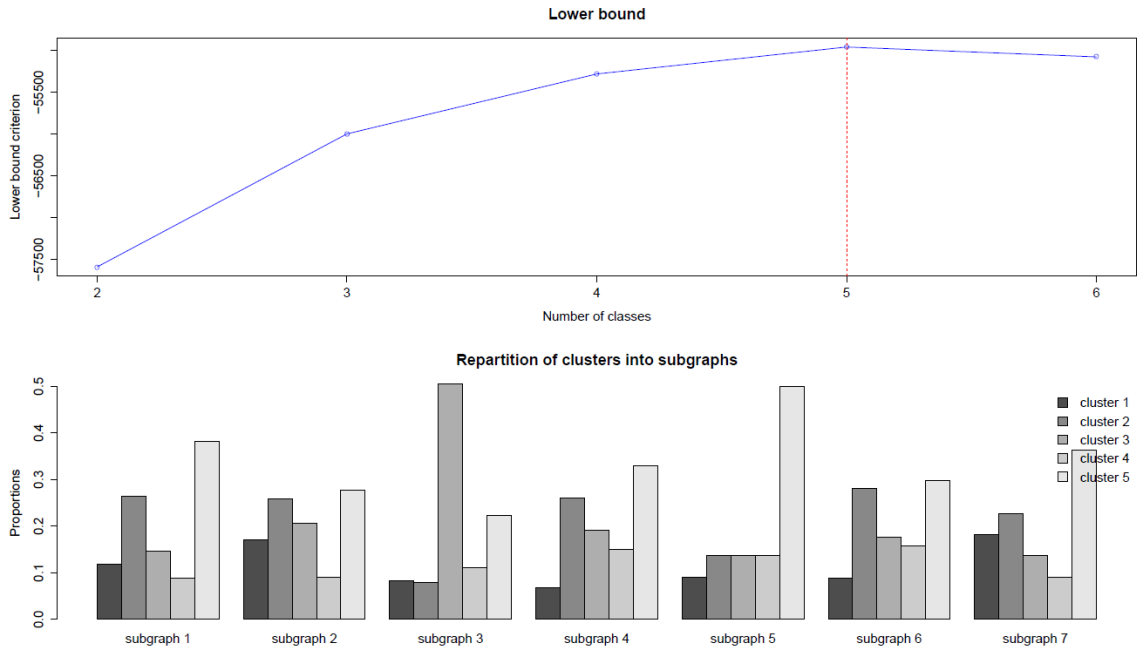

Figure 12.7: Outputs of the rsm function: values of the model selection criterion for the different values of K (top) and proportions of the latent clusters in the 7 subgraphs (bottom). The subgraphs are here associated to continents

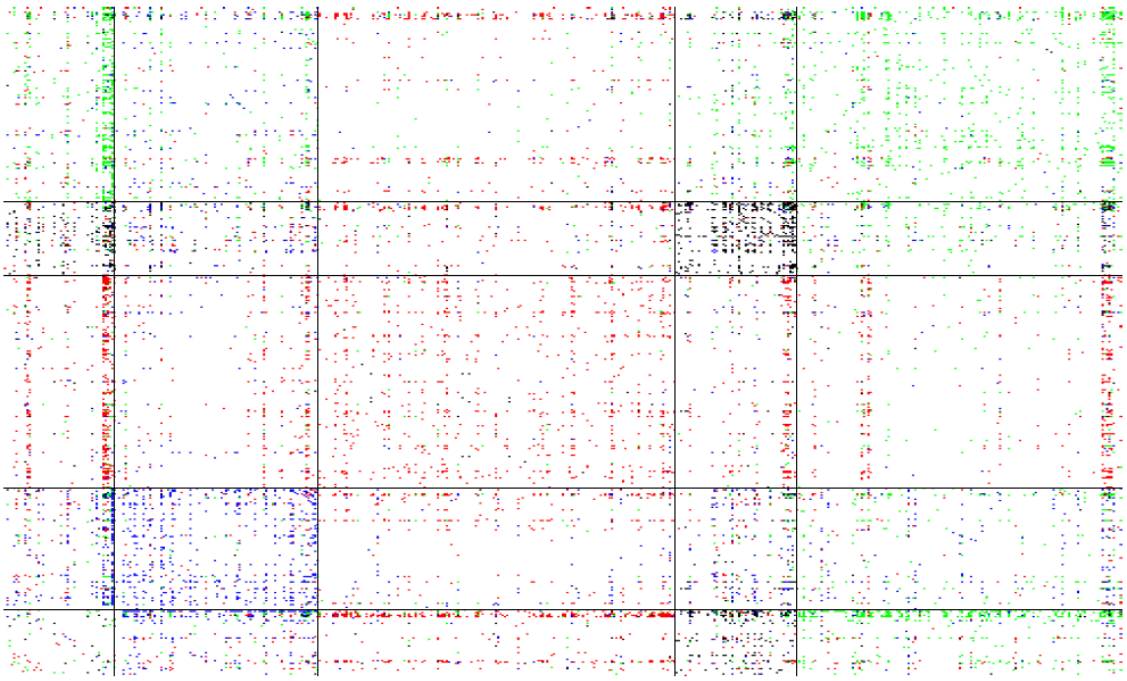

Figure 12.8: Reorganised adjacency matrix according to the partition into 5 latent groups found by Rambo: containers (black), solid bulk (red), liquid bulk (green) and passengers (blue)

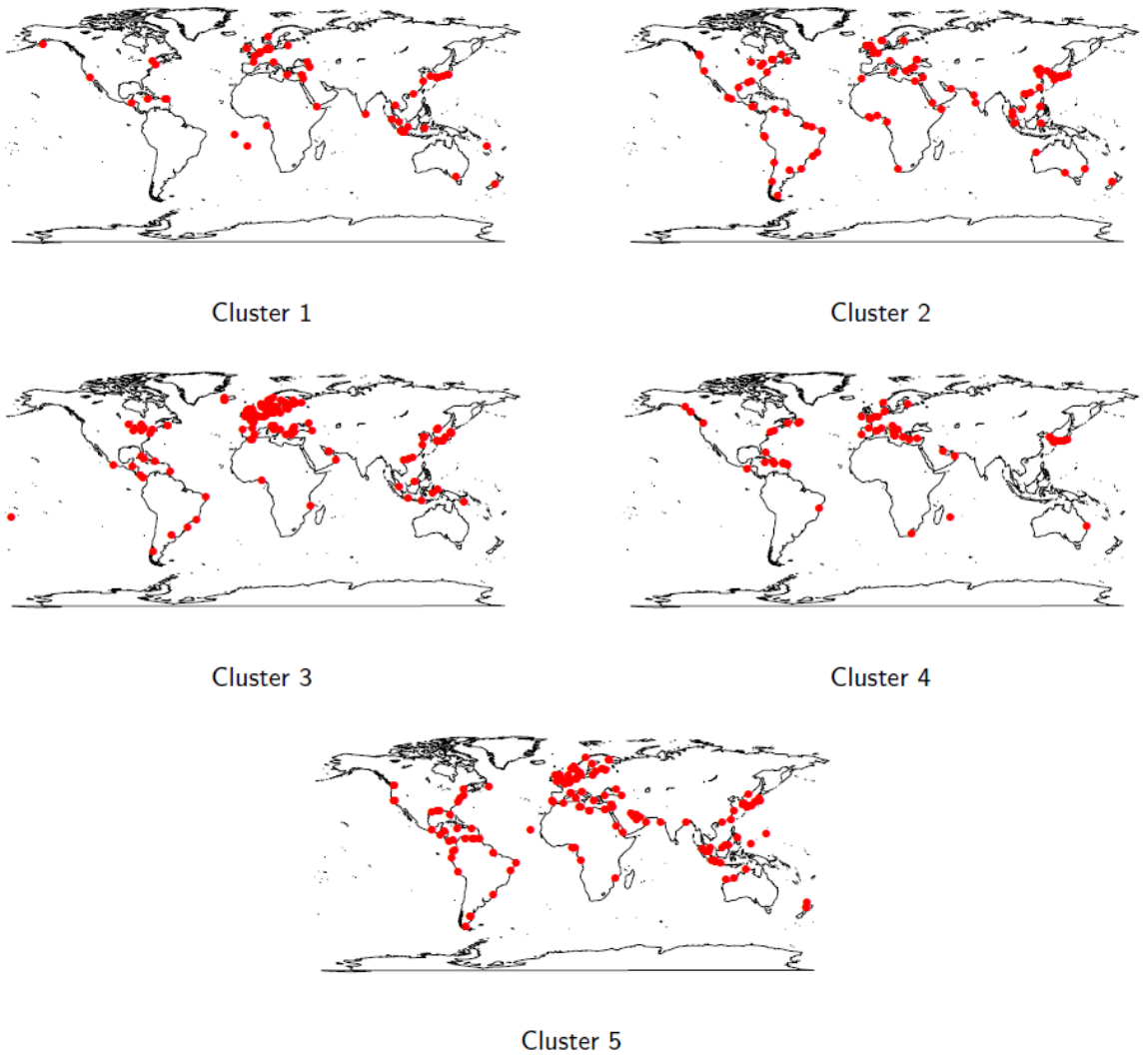

Figure 12.9: Geographical localization of the 5 clusters found by Rambo for the RSM model