O SELF-ORGANIZING MAP COMO FERRAMENTA NA ANÁLISE GEO- DEMOGRÁFICA

Miguel Loureiro, Fernando Bação, Instituto Superior de Estatística e Gestão de Informação Universidade Nova de Lisboa, Campus de Campolide 1070-312 Lisboa
Fax: 21 387 21 40, e-mail:{mloureiro,bacao}@isegi.unl.pt
Palavras Chave: geo-demografia, self-organizing map, classificação, Lisboa

1. Resumo

Este trabalho tem como principal objectivo a exploração do Self-Organizing Map (SOM) como ferramenta para a análise geo-demográfica, no caso particular dos dados referentes à Área Metropolitana de Lisboa (AML). Paralelamente, e como forma de estabelecer um termo de comparação para a performance do SOM (Kohonen *et all*, 1995), desenvolveu-se uma breve análise com base no cálculo de índices demográficos tradicionais. Em ambos os casos, SOM e índices demográficos, os resultados foram mapeados por forma a permitir uma avaliação geograficamente contextualizada e retirar algumas conclusões sobre a valia de ambos os métodos.

2. Introdução

A existência de grandes bases de dados é comum hoje em dia nas mais diversas áreas, desde o comércio, à indústria, passando pela investigação, por razões que se prendem com a evolução das tecnologias de recolha de informação. A evolução dos computadores, dos métodos de aquisição de dados, a generalização de pagamentos automáticos, entre muitos outros factores, criaram uma enorme riqueza de dados que é necessário transformar em informação. Novas ferramentas que permitam a análise de grandes volumes de dados tornam-se, portanto, necessárias, por forma a gerar conhecimento a partir dessa informação.

No caso concreto da análise geo-demográfica de regiões, a quantidade, multidimensionalidade e variância dos dados é frequente, pelo que a sua análise não é trivial. É, portanto, uma situação que justifica uma abordagem típica de *Data Mining*, como por exemplo a utilização de redes neuronais como o SOM.

A utilização de índices demográficos tradicionais neste tipo de análise tem por vezes interpretações difíceis, devido ao número de índices calculados e à necessidade de visualizar cada um deles em mapas separados. Por outro lado, a utilização de ferramentas que efectuem uma análise conjunta de todos os dados, como é o caso do SOM, vêm permitir uma maior facilidade na interpretação dos mesmos. Nesse sentido, o objectivo deste artigo é confrontar estes dois métodos de análise para o conjunto de dados da AML, ao nível da secção estatística.

No capítulo 3 é apresentada a metodologia de análise dos dados, no capítulo 4 são apresentados os resultados dos métodos acima referidos e as conclusões estão apresentadas no capítulo 5.

3. Metodologia

3.1. Análise exploratória de dados

Os dados utilizados neste trabalho consistem numa tabela de 3968 registos, um para cada secção estatística da AML, com dezanove atributos para cada um. Os dados resumem-se em três grupos: um campo, a chave primária, que identifica univocamente cada secção estatística; dois campos que representam as coordenadas x e y dos centróides da área de cada secção estatística; e os restantes dezasseis campos, todos referentes a grupos etários de homens e de mulheres.

O primeiro passo deste trabalho consistiu numa análise exploratória dos dados. Desta análise pode afirmar-se que a maioria dos campos tem *outliers*. Os dezasseis campos referentes aos grupos etários têm uma ordem de grandeza próxima, ao contrário dos dados dos campos referentes às coordenadas dos centróides das secções estatísticas, que têm uma ordem de grandeza completamente diferente. Estas conclusões iniciais foram importantes para o decorrer do trabalho, nomeadamente na preparação dos dados anterior ao treino do SOM.

3.2. Análise a partir de índices demográficos

Numerosos índices demográficos são habitualmente calculados em análise geo-demográfica. No presente trabalho foram calculados três dos principais índices demográficos para cada secção estatística, visando a comparação entre três grupos etários distintos: população jovem (dos zero aos catorze anos), população activa (dos quinze aos sessenta e quatro anos) e população idosa (com mais de sessenta e cinco anos).

A escolha destes três grupos etários permitiu a redução da dimensionalidade do espaço de *input* de dezasseis para três dimensões, facilitando assim a análise, mas provocando uma perda de informação significativa.

Os índices calculados foram os apresentados abaixo, e relacionam os três grupos etários entre si.

$$\textit{Índice de envelhecimento} = \frac{\textit{População} + 65\textit{anos}}{\textit{População}\ 0-14\textit{anos}}\ x100$$

$$\textit{Índice de dependência dos idosos} = \frac{\textit{População} + 65\textit{anos}}{\textit{População}\ 15-64\textit{anos}}\ x100$$

$$\textit{Índice de dependência dos jovens} = \frac{\textit{População}\ 0-14\textit{anos}}{\textit{População}\ 15-64\textit{anos}}\ x100$$

Os resultados para cada secção estatística foram posteriormente mapeados com através de um sistema de informação geográfica.

3.3. Análise com um SOM

O SOM pode ser caracterizado como uma rede neuronal não supervisionada, que pode ser encarada como uma projecção não linear de dados multidimensionais, estando por esta razão completamente livre para se ajustar aos dados de *input*. Por forma a processar os dados há que começar por ajustar os parâmetros, como o raio de vizinhança topológica, o número de neurónios, e a taxa de aprendizagem. As possibilidades de análise que os

outputs fornecidos permitem, entre as quais as matrizes U (Vesanto, 1999), constituem ferramentas de inestimável valor na compreensão da estrutura interna dos dados. Assim sendo, o SOM foi utilizado para agrupar os indivíduos em estudo, as secções estatísticas da AML, em *clusters*, que posteriormente foram mapeados através de um sistema de informação geográfica.

Após a análise exploratória dos dados atrás referida, o procedimento de análise com o SOM seguiu os seguintes passos: preparação dos dados, análise de sensibilidade aos parâmetros do SOM, treino do SOM, visualização dos resultados, agregação inicial de indivíduos após análise da matriz U, agregação posterior de indivíduos dentro de cada grupo inicial treinando o SOM com os respectivos indivíduos. No final, determinaram-se os valores médios dos *clusters*, sendo estes parâmetros essenciais na interpretação do significado de cada grupo.

O primeiro passo na preparação dos dados foi a normalização dos mesmos. Verificou-se na análise exploratória dos dados que os vários campos tinham ordens de grandeza e unidades diferentes, pelo que normalizar os dados tornou-se imperativo. Outra razão que obrigou a normalização deve-se ao facto que a maioria dos modelos não paramétricos, aqueles que dependem essencialmente do uso das distâncias entre os dados como o caso do SOM, assumirem que as diferentes direcções do espaço de *input* possuem o mesmo peso (Bação, 2004). A normalização escolhida foi a z-score, na qual cada variável de *input* é transformada por forma a ter média igual a zero e desvio padrão igual a um, por duas razões: verificou-se a existência de *outliers* na maioria dos campos, o que inviabilizava a normalização min-max; e os *outliers* não tinham valores extremos, não justificando por isso a utilização de uma normalização sigmoidal.

Coligiram-se os dados em 2 grupos distintos, incluindo ou não os campos correspondentes às coordenadas dos centróides de cada secção estatística. Para cada um dos grupos de dados foram efectuados todos os passos do estudo, o que permitiu verificar a influência da localização geográfica na constituição dos *clusters* das referidas unidades territoriais.

A análise de sensibilidade aos parâmetros do SOM consistiu em efectuar numerosos testes à influência que os seus parâmetros têm ajuste da rede aos dados. Variaram-se o número de épocas, de taxa de aprendizagem e de raio de vizinhança topológica, para um dado mapa topológico, por forma a se perceber a influência dos mesmos no resultado final. O indicador que permitiu verificar a qualidade do ajuste da rede neuronal aos dados foi o erro de quantização, que mede o somatório das distâncias de cada indivíduo ao neurónio mais próximo.

Uma vez percebida a influência de cada parâmetro, procedeu-se ao treino do SOM. Foi escolhida uma forma hexagonal do mapa topológico, em detrimento de uma forma rectangular, por facilitar a visualização; e a dimensão de trinta e seis neurónios, por ser substancialmente inferior ao número de indivíduos, mas ainda assim permitir uma boa segmentação dos mesmos e toda a variabilidade no espaço multidimensional de *input*.

Inicialmente no treino da rede, os vectores no espaço multidimensional de *input* correspondentes aos neurónios são definidos aleatoriamente. O treino prossegue após definição dos parâmetros número de épocas, taxa de aprendizagem e raio de vizinhança topológica. Este passo é efectuado pelo menos duas vezes, pela necessidade de refinar a aprendizagem, utilizando no final taxas de aprendizagem e raios de vizinhança topológica menores, e número de épocas maior.

A análise da matriz U do treino com menor erro de quantização permitiu a agregação dos neurónios mais próximos entre si em *clusters*. A maior ou menor distância entre neurónios é fornecida por uma escala de cinzentos. A sua interpretação contém alguma subjectividade, estando esta análise, por essa razão, fortemente dependente da experiência do utilizador.

Com base na correspondência indivíduo/neurónio mais próximo fornecida pelo software, foi possível determinar quais os indivíduos agregados entre si. Nesta constituição dos *clusters*, verificou-se que os indivíduos de um dos grupos apresentavam uma distância considerável inter e intra grupo. Por essa razão, optou-se por repetir todo o processo de treino e de análise para os referidos indivíduos, por forma a identificar *sub-clusters*

dentro do *cluster* inicial em apreço. Uma vez identificados todos os grupos de indivíduos, o seu mapeamento foi efectuado com recurso a um sistema de informação geográfica.

Como referido anteriormente, a média de cada grupo foi calculada para todos os campos, por forma a ser possível a interpretação dos resultados.

4. Apresentação e discussão de resultados

4.1. Índices demográficos

Nas figuras 1, 2 e 3 apresentam-se os resultados dos três índices demográficos calculados, mapeados para as secções estatísticas da AML.

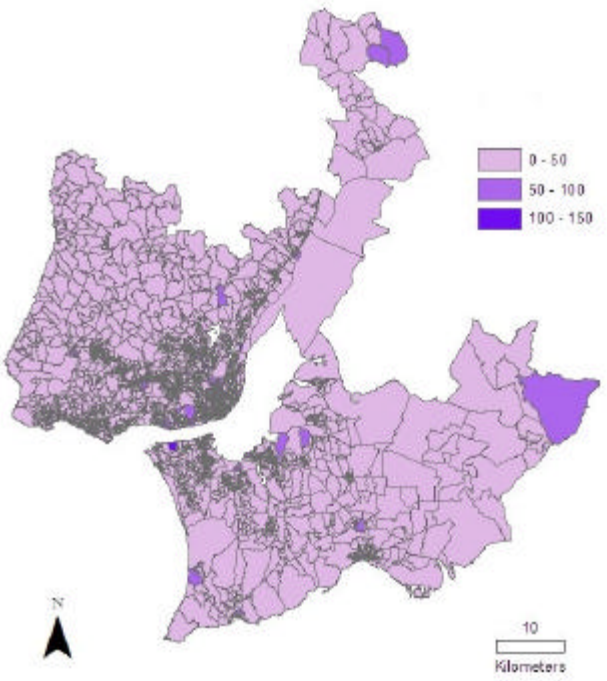


Figura 1 – Índice de dependência dos idosos

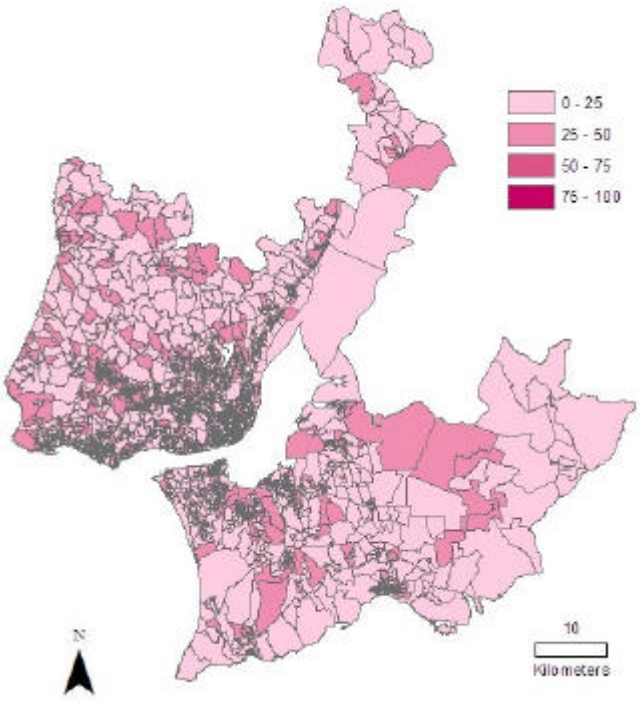


Figura 2 – Índice de dependência dos jovens

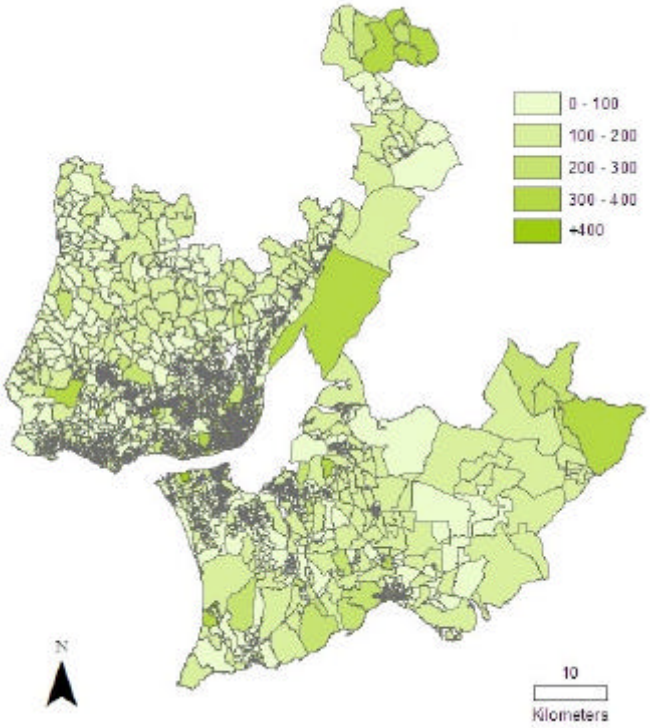
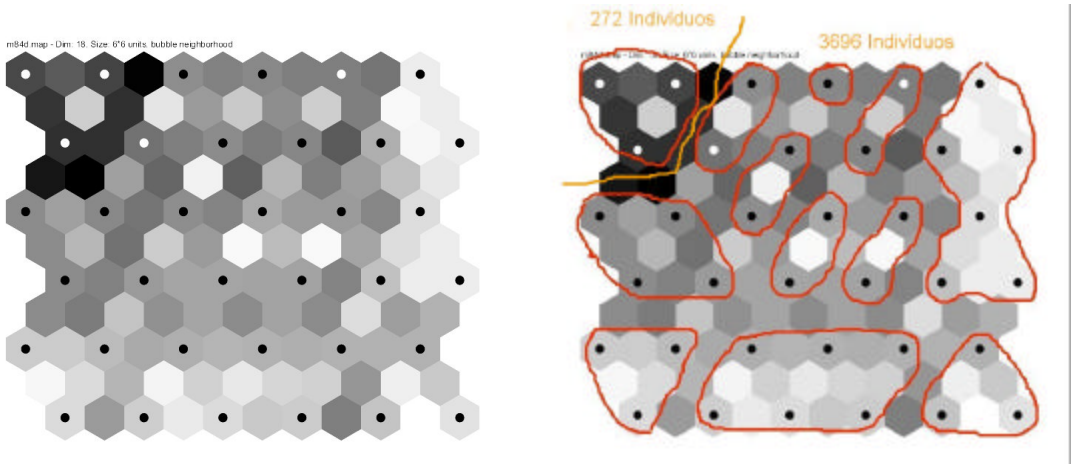


Figura 3 – Índice de envelhecimento

Verifica-se que a análise a partir deste método é difícil, dado a necessidade de interpretar vários mapas. Adicionalmente a interpretação de cada um deles está fortemente dependente da escala utilizada, pelo que as conclusões poderão ser enviesadas por este parâmetro. Por outro lado verifica-se que a classificação de cada secção estatística é independente das restantes, o que permite comparações precisas ao longo do espaço e do tempo.

4.2. SOM

Apresentam-se na Figura 4 as matrizes U do primeiro treino do SOM, referentes aos dados que incluem os centróides das secções estatísticas.



a) b)

Figura 4 – Matrizes U após primeiro treino do SOM para dados com centróides, a) antes de agrupamento de neurónios, b) após agrupamento de neurónios

O agrupamento de neurónios apresentado na Figura 4 b) foi efectuado com base na respectiva escala de cinzentos. Cor clara entre os neurónios representa que estes estão próximos entre si, cor escura representa o oposto. Assim sendo, os neurónios agrupados tendem a ter fronteiras claras entre si. Todavia, podem ocorrer situações nas quais o agrupamento é dúbio, devido a factores como a variância dos dados, originado grupos menos bem definidos. Esta análise visual, carece de experiência por parte do utilizador, e é subjectiva, pelo que outro agrupamento de neurónios seria passível de ser efectuado. No

entanto, é relativamente óbvio para qualquer utilizador experimentado a identificação das principais características presentes nos dados.

Durante a análise da matriz U verificou-se que existem três neurónios bastante distantes dos restantes, sendo claramente identificados pelas fronteiras a negro na matriz (zona designada 1 a 4 representados na Figura 5) . Foi efectuado um segundo treino da rede aos 272 indivíduos que lhes correspondiam, tendo originado um agrupamento dos mesmos em quatro *clusters*. Assim sendo agruparam-se o total de indivíduos em quinze *clusters*, como apresentado na Figura 5.

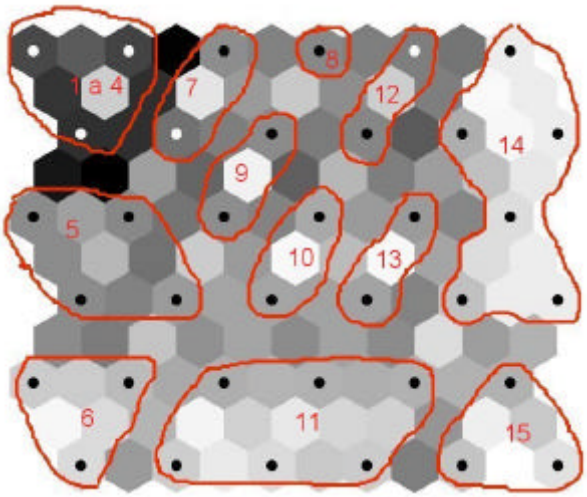


Figura 5 – Identificação das características de cada *cluster*, dados com centróides

Como exemplo do tipo de análise que a matriz U permite optámos por analisar um *cluster* em particular (*cluster* 14 da Figura 5). Na Figura 6 está representado o referido *cluster* sendo este caracterizado por secções estatísticas cuja população é bastante idosa e cujo número de habitantes é muito abaixo da média. Em termos genéricos poderíamos designar este *cluster* como "zonas populacionalmente desertificadas", este facto fica-se a dever, na maioria dos casos, ao envelhecimento populacional. Existem também casos de grandes secções pouco povoadas, essencialmente devido a restrições relacionadas com áreas protegidas.

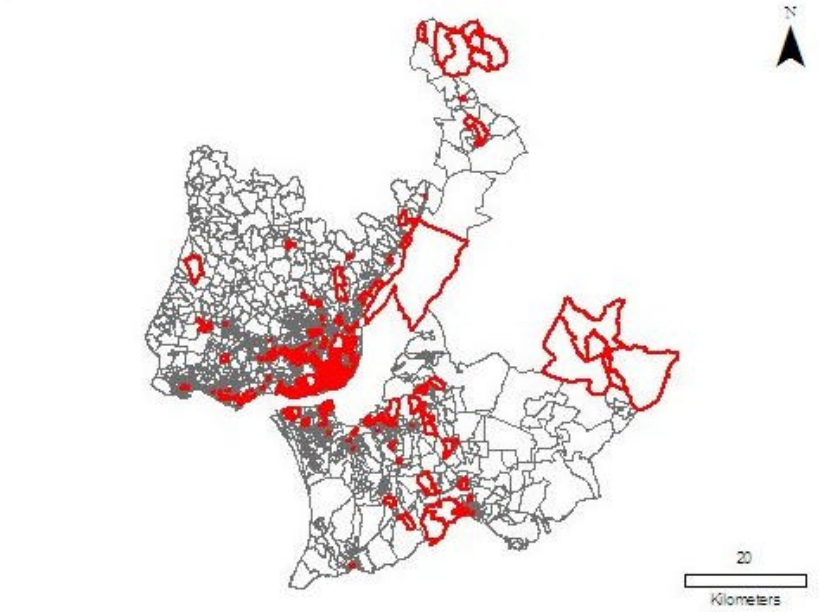


Figura 6 – Mapeamento do *cluster* 14, dados com centróides

A tendência de cada variável pode facilmente ser observada no respectivo plano, outro dos outputs do *software*. Na Figura 7 apresenta-se o plano da primeira variável do estudo, correspondente ao grupo etário dos zero aos quatro anos de idade. Neste caso facilmente concluímos que os *clusters* onde o grupo etário dos indivíduos com 4 ou menos anos se concentra no topo esquerdo da matriz. A utilização dos planos das variáveis é particularmente útil na medida em que permite uma rápida interpretação da forma como as diferentes variáveis variam ao longo da matriz.

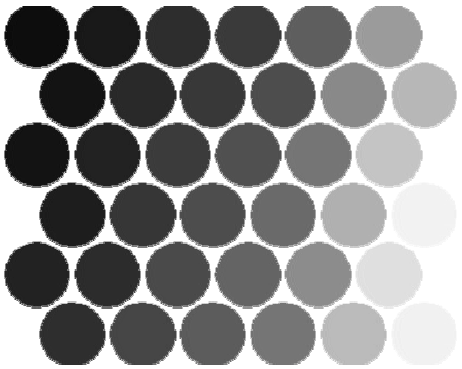


Figura 7 – Plano da variável 1 (indivíduos com 4 ou menos anos)

O agrupamento de indivíduos atrás referido permitiu, com auxílio de um sistema de informação geográfica, obter os mapas das figuras 8 a 11.

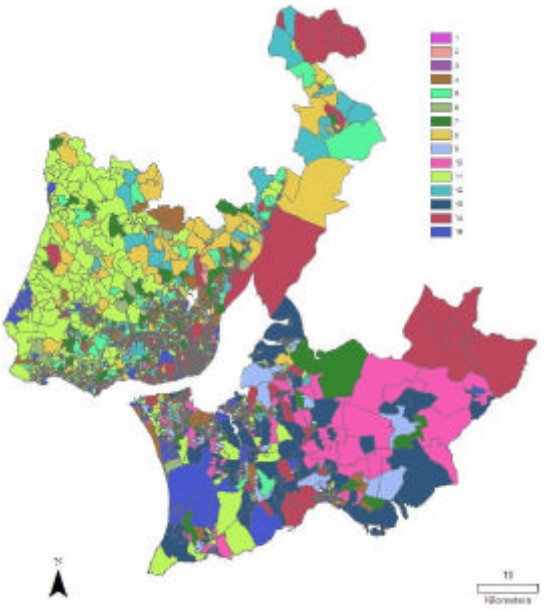


Figura 8 – Mapeamento dos *cluster*s na AML, dados com os centróides das secções estatísticas

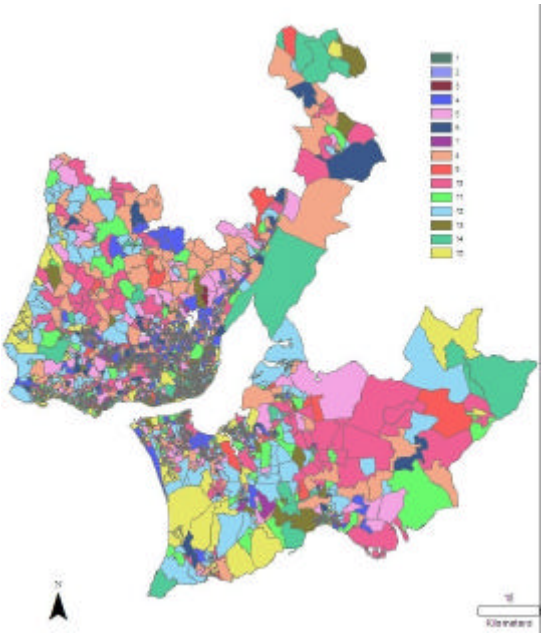


Figura 9 – Mapeamento dos *clusters* na AML, dados sem os centróides das secções estatísticas

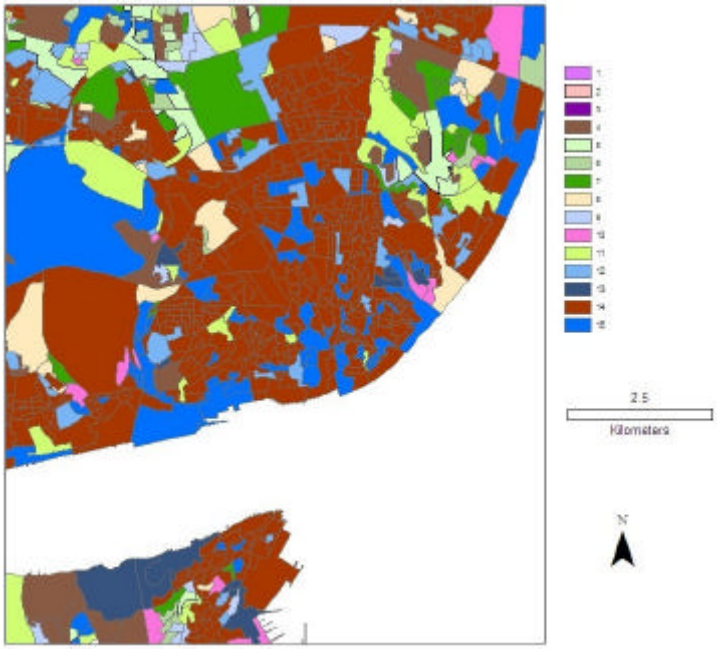


Figura 10 – Mapeamento dos *clusters* na zona do centro de Lisboa, dados com os centróides das secções estatísticas

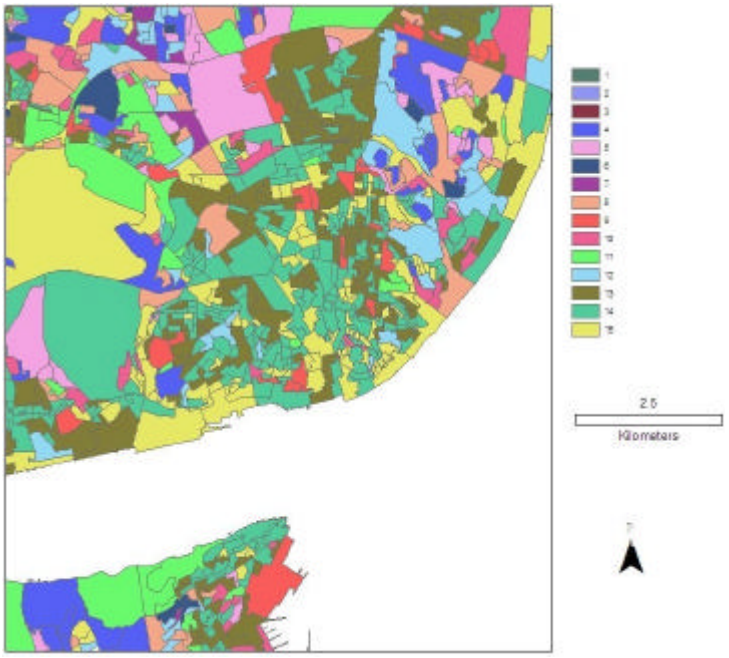


Figura 11 – Mapeamento dos *clusters* na zona do centro de Lisboa, dados sem os centróides das secções estatísticas

Verificou-se que a inclusão das coordenadas dos centróides teve um efeito aglutinador de secções próximas entre si, produzindo mapas mais homogéneos. Note-se que o número de *clusters* não teria necessariamente de ser igual nos dois estudos, apenas após análise das respectivas matrizes U se chegou a este resultado.

Como últimas notas torna-se pertinente referir que a análise com o SOM é relativa aos indivíduos que forem incluídos no estudo. Este poderá ser um ponto fraco desta abordagem visto ser de difícil comparação caso se decidisse comparar resultados com outra área geográfica que não a AML. Por outras palavras, ao contrário dos índices demográficos que possuem um significado próprio, e independente do contexto de aplicação, o SOM produz classificações relativas onde o significado esta dependente do contexto em que é aplicado.

5. Conclusões

Uma das principais conclusões deste trabalho relaciona-se com o potencial do SOM na análise de grandes quantidades de dados caracterizados por elevada dimensionalidade. Por outro lado, a utilização dos índices demográficos revelou algumas limitações na medida em que não permite construir uma "imagem" global das principais características do conjunto de dados. Assim sendo torna-se necessário a apresentação de um mapa por cada índice calculado, cuja interpretação fica fortemente dependente da escala utilizada em cada um.

Uma das vantagens da análise com índices demográficos relaciona-se com o facto de ser intemporal e invariável no espaço, isto é, cada índice permite uma comparação absoluta entre secções estatísticas. Por outro lado, a análise com o SOM é relativa aos indivíduos que forem incluídos no estudo.

Adicionalmente, conclui-se que a análise das matrizes U, sendo decisiva na qualidade das conclusões, contém em si mesma alguma subjectividade, sendo necessária experiência por parte do utilizador por forma a garantir resultados fiáveis. Adicionalmente, os planos das variáveis constituem um precioso auxiliar à análise permitindo uma rápida identificação das principais tendências das variáveis utilizadas.

O efeito da inclusão dos campos correspondentes às coordenadas dos centróides das secções estatísticas teve um efeito aglutinador das secções próximas entre si, resultado que veio confirmar as expectativas.

Finalmente, podemos dizer que de forma genérica os resultados da análise com o SOM também correspondem às expectativas, apresentando um núcleo central da cidade composto por população envelhecida e as novas zonas de crescimento com uma estrutura etária mais equilibrada.

6. Referências

Bação, F. (2004). Apontamentos da disciplina de Data Mining Geo-Espacial: ISEGI - UNL.

Kohonen, T., Hynninen, J., Kangas, J., & Laaksonen, J. (1995). SOM_PAK - The Self-Organizing Map Program Package (Version 3.1).

Nazareth, J. M. (2000). *Introdução à Demografia*. Lisboa: Editorial Presença.

Vesanto, J. (1999). *SOM-Based Data Visualization Methods.* Helsinki University of Technology, Helsinki.

Vesanto, J. (2000). *Using SOM in Data Mining.* Unpublished Licentiate, Helsinki University of Technology, Helsinki.

Vesanto, J., & Alhoniemi, E. (2000). Clustering of the Self-Organizing Map. *IEEE TRANSACTIONS ON NEURAL NETWORKS, II*(3).