Finisterra, XXXVIII, 75, 2003, pp. 135-147

ASPECTOS METODOLÓGICOS DA UTILIZAÇÃO DO *DATA MINING* NO ÂMBITO DA GEOGRAFIA

FERNANDO LUCAS BAÇÃO [1]
MARCO PAINHO [2]

Resumo – Depois de, nas últimas décadas do século passado, ter sofrido transformações importantes, decorrentes em larga medida dos desenvolvimentos da Ciência Computacional, a Geografia encontra-se perante mais uma oportunidade, o *Data Mining* Geo-espacial. O *Data Mining* representa uma nova geração de ferramentas, incontornável para todos os que procuram dominar a complexidade, independentemente da área de trabalho. Baseia-se na riqueza das actuais bases de dados e na capacidade de processamento computacional, para implementar uma abordagem eminentemente indutiva, capaz de resolver problemas muito para além do actual estado do conhecimento teórico. No entanto, para que possa contribuir para o desenvolvimento teórico é indispensável adoptar um quadro metodológico adequado, que permita escapar a uma lógica puramente indutiva. Neste contexto avançamos com a proposta de uma epistemologia *popperiana* onde a actividade conjectural do investigador constitui o cerne do trabalho científico.

Palavras-chave : *Data Mining*, *Data Mining* geo-espacial, SIG, geografia.

Abstract – METHODOLOGICAL ASPECTS OF USING *DATA MINING* IN GEOGRAPHICAL WORKS. At the start of the new century, Geography finds itself in a very interesting position. After various developments made towards the end of the twentieth century mainly due to the rapid progress in the field of computation, geographers now face a new challenge: geospatial data mining. Geospatial data mining could be the key with which geography achieves the scientific objectives at the basis of the quantitative revolution.

It is true that most of the opportunities that have enabled geography to grow have originated outside the discipline; it is also true that geographers have always had the intelligence and vision to rapidly adopt those innovations in the practical work of their profession. The quantitative revolution and Geographical Information Systems are important steps in the search for knowledge about space and spatial-temporal patterns, which are of enormous complexity.

A new generation of tools has emerged and grown indispensable to anyone concerned with mastering complexity in any field of knowledge. These tools make use of the huge resources and dimension of today's databases and have ever-growing

[1] ISEGI-UNL - Campus de Campolide, 1070-124 Lisboa. Tel: +351 213 870 413 – Fax: +351 213 872 140. E-mail: bacao@isegi.unl.pt

[2] ISEGI-UNL - Campus de Campolide, 1070-124 Lisboa. Tel: +351 213 870 413 – Fax: +351 213 872 140. E-mail: painho@isegi.unl.pt

processing power as their main ally in establishing an inductive approach capable of solvingproblems far beyond the reach of theoretical formulations. Nevertheless, in order for these tools to contribute to the essential theoretical development of geography, it is necessary to adopt an appropriate methodological framework and escape pure inductive logic. Accordingly, a Popperian epistemological framework is proposed, in which the conjectural activity of the researcher is the core of the scientific work.

Key words : Data Mining, geo-spatial Data Mining, GIS, geography.

Resumé – Aspects méthodologiques de l'utilisation géographique du *data mining*. Après les transformations importantes qu'elle a subie au cours des dernières décennies du XX^e^ siècle grâce, en bonne partie, au progrès des Sciences télé-informatiques, la Géographie se trouve face à une nouvelle opportunité, le *Data Mining Geo-spatial*. C'est une nouvelle génération d'instruments, qui permet de dominer la complexité des faits, quel que soit le domaine de travail. Or, on sait que la Géographie a toujours progressé en absorbant des techniques venues d'autres branches scientifiques.

Grâce à la richesse actuelle des bases de données et à l'énorme capacité de leur manipulation par ordinateur, on parvient à développer une approche essentiellement inductive, capable résoudre des problèmes encore inabordables à partir des actuelles connaissance théoriques. Pourtant, afin de contribuer au progrès théorique, il serait indispensable d'adopter un cadre méthodologique adapté, qui permette d'échapper à une logique purement inductive. C'est pourquoi on propose une nouvelle épistémologie *popperienne*, dans laquelle l'activité conjoncturelle du chercheur constituerait le noyau même du travail scientifique.

Mots-clés : *Data Mining*, *Data Mining* geo-spatial, SIG, géographie.

I. INTRODUÇÃO

Retrospectivamente, podemos com maior facilidade encaixar as peças do *puzzle* que caracterizam a evolução da Geografia nas últimas décadas do século XX. A Revolução Quantitativa, mais do que uma verdadeira revolução, constituiu por um lado o reconhecimento de que era necessária uma abordagem mais científica, traduzindo uma visão positivista da Geografia (Johnston, 1986; Bird, 1993); por outro lado foi capaz de iniciar o desenvolvimento de um quadro conceptual organizado e consistente (Macmillan, 1997; Couclelis, 1998). Os objectivos desta revolução esbarraram numa insuficiência instrumental, quer em termos tecnológicos, quer em termos metodológicos (Openshaw, 1993). Os Sistemas de Informação Geográfica (SIG) resolveram, em larga medida, a primeira parte do problema, proporcionando uma sofisticada plataforma tecnológica. No entanto, a segunda parte do problema permanece por resolver (Rogerson *et al.*, 1994; Openshaw, 1991, 1993, 1999; Goodchild, 1991; Bailey, 1994). As tentativas para solucionar este problema consistiram, essencialmente, na «importação» dos métodos da estatística clássica para o domínio da análise dos dados geo-referenciados (Rogerson *et al.*, 1994; O'Kelly *et al.*, 1994; Getis, 1994; Abler *et al.*, 1977). Esta abordagem não conheceu o sucesso desejado devido a duas causas essenciais. Por um lado, a utilização destes métodos exige um conhecimento teórico sobre o domínio de aplicação (Kennedy *et al.*, 1998; Hand, 1999) que ainda não existe na Geografia (Rogerson *et al.*, 1994; Openshaw, 1991). Por outro lado, os dados geo-referenciados possuem particularidades que inviabilizam a utilização da maioria dos métodos estatísticos (Anselin, 1989; Bailey, 1994; Gahegan, 2001;

Openshaw, 1993). O desafio consiste em resolver a segunda parte do problema. Cumprido este desígnio talvez se assista à verdadeira Revolução.

O *Data Mining* (DM) (Han *et al.*, 2001; Weiss *et al.*, 1998), também designado *Descoberta de Conhecimento*, constitui uma nova oportunidade para resolver o constrangimento metodológico que tem impedido a Geografia e a Ciência da Informação Geográfica (CIG) de cumprir todo o seu potencial (Gahegan, 2001; Openshaw, 1999). Existem, no entanto, alguns problemas associados à utilização destas novas ferramentas no âmbito da Geografia, que é necessário ter em conta por forma a garantir um avanço efectivo do conhecimento que hoje dispomos sobre os fenómenos espaciais. É sobre estas questões que reflectiremos neste artigo. Assim, a secção II será dedicada à definição de DM e suas principais características. A secção III centra-se na relação entre o DM e a Geografia, bem como nas características que definem a actual informação geográfica. Na secção IV explicitamos a nossa perspectiva sobre o quadro metodológico em que o DM pode ser utilizado no desenvolvimento teórico, por oposição às abordagens puramente indutivas. Finalizamos com algumas conclusões sobre as consequências da adopção dos quadros conceptuais apresentados.

II. *DATA MINING*

Como noutros campos do conhecimento, também no DM existem diversas definições, diversas perspectivas e opiniões. Aqui a questão da definição é um pouco mais complicada, na medida em que esta área do conhecimento cresceu com contribuições muito diversas (um pouco à semelhança do que aconteceu com os Sistemas de Informação Geográfica). Na génese e desenvolvimento do DM encontramos diferentes áreas de investigação, como a Estatística, a Inteligência Artificial/Reconhecimento de Padrões, a Ciência Computacional, entre outros. Como seria de esperar, investigadores de proveniências diferentes abordam a temática de forma diferente e com interesses diversos, o que determina perspectivas diferentes sobre o que é o DM. Genericamente, poderemos dizer que o DM é *o processo não trivial de identificar nos dados padrões que sejam válidos, potencialmente úteis e compreensíveis* (Fayyad *et al.*, 1996), procurando traduzir dados em informação, informação em conhecimento, que por sua vez proporciona a oportunidade de agir com propriedade e racionalidade.

Uma das marcas distintivas do DM relaciona-se com a intensividade computacional dos algoritmos utilizados, que desta forma tendem a substituir a elegância formal dos métodos matematicamente consubstanciados, pela busca baseada em potentes algoritmos de optimização. Assim, o DM constitui um dos principais beneficiários dos constantes aumentos de capacidade de processamento e muito especialmente da possibilidade de recorrer às arquitecturas de processamento paralelo.

Um outro factor particularmente relevante, para enquadrar o DM, tem que ver com as características das actuais bases de dados e que se traduz em dois aspectos essenciais. Por um lado, a maior dimensionalidade das bases de dados hoje disponíveis, que se traduz na enorme quantidade de campos utilizados para descrever cada registo. O DM permite aos utilizadores uma exploração mais efectiva deste «espaço de *input*», ou seja, permite a possibilidade de explorar de forma mais efectiva toda a dimensionalidade disponível nas bases de dados. Por outro lado, amostras de maior dimensão produzem erros de estimação e variâncias mais pequenos, permitindo aos utilizadores inferências seguras sobre segmentos relativamente pequenos da população. Não raras vezes a amostra coincide com o próprio universo, sendo que, neste caso, os testes de significância deixam de

ter sentido, na medida em que o valor observado da estatística coincide com o valor do parâmetro (HAND, 1999).

Uma outra forma, talvez mais útil no contexto deste artigo, de proceder à definição de DM consiste em recorrer à enumeração das ferramentas utilizadas. As ferramentas mais emblemáticas do DM são:

- **Algoritmos Genéticos**: técnicas de optimização que imitam os processos biológicos subjacentes à teoria *darwinista* da evolução das espécies (um exemplo de aplicação no âmbito de problemas geográficos pode ser encontrado em BAÇÃO *et al.*, 2002);
- ***Clustering***: conjunto de técnicas que permitem proceder à partição de um conjunto de dados (ou objectos) em grupos, sendo que objectos semelhantes ficam no mesmo grupo (*cluster*);
- **Regras de Associação**: consiste na extracção de regras *if-then* com propriedades de significância estatística a partir dos dados;
- **Método do vizinho mais próximo** (*nearest neighbour*): é uma técnica que classifica cada registo numa base de dados, baseado na combinação das classes dos *k* registos mais próximos;
- **Visualização**: traduz-se na interpretação visual de relações complexas em conjuntos de dados multidimensionais;
- **Redes Neuronais Artificiais**: modelos preditivos não-lineares que «aprendem» através do treino e são inspirados nas redes neuronais biológicas, como no cérebro humano;
- **Árvores de Decisão**: estruturas em forma de árvore que representam conjuntos de decisões. Estas decisões geram regras para a classificação de conjuntos de dados. Dois dos métodos mais conhecidos são o *Classification and Regression Trees* (CART) e o *Chi Square Automatic Interaction Detection* (CHAID). Ambos produzem um conjunto de regras que podem ser aplicadas a um novo (não-classificado) conjunto de dados por forma a prever quais os registos que terão um determinado resultado.

Por forma a perspectivar com maior rigor o tipo de contribuição que o DM poderá trazer à Geografia vale a pena enquadrar os tipos de modelos que servem de base à ciência. Assim, começamos por propor uma divisão dos modelos utilizados na ciência em três tipos fundamentais (KENNEDY *et al.*, 1998):

- modelos determinísticos;
- modelos paramétricos;
- modelos não-paramétricos.

Tal como em outras tarefas levadas a cabo na Informática, a modelação requer um «programa» que proporciona instruções detalhadas sobre a forma como o processo se desenvolverá. Estas instruções são tipicamente equações matemáticas, que caracterizam a relação existente entre *inputs* e *outputs*. A formulação destas equações constitui o problema central da modelação. A melhor forma de modelar consiste em formular equações «fechadas» que definem deterministicamente a forma como os *outputs* são obtidos a partir dos *inputs*. Sendo todas as características constantes referimo-nos a eles como modelos determinísticos. Este tipo de modelo é apropriado para o tratamento de problemas simples e perfeitamente compreendidos. Infelizmente, a maioria dos proble-

mas, especialmente nas ciências sociais e humanas, não se prestam a uma descrição tão simples, ou não são tão bem conhecidos como os que caracterizam a Física Newtoniana.

No caso dos modelos paramétricos enfrentamos um problema de estimação. Apesar de possuir uma boa ideia sobre a forma como *inputs* e *outputs* se relacionam, não possuímos o grau de certeza necessário de um modelo determinístico. Uma forma de ultrapassar o problema da ausência de conhecimento consiste em substitui-lo por experimentação. Assim, através da experimentação estimamos os parâmetros em falta, por forma a que o modelo produza estimativas tão próximas da realidade quanto possível. O aspecto fundamental dos modelos paramétricos consiste no facto de equações matemáticas explícitas caracterizarem a estrutura da relação entre *inputs* e *outputs*, havendo, no entanto, alguns parâmetros não especificados. Estes são escolhidos através da análise dos exemplos de que dispomos, por outras palavras, são estimados a partir de uma amostra. A regressão linear constitui uma das aplicações mais conhecidas de modelação paramétrica que assume como hipótese a existência de uma relação linear entre *inputs* e *output*.

Existem ainda os modelos não-paramétricos, ou seja o tipo de modelo normalmente utilizado no DM, que podemos definir como modelos que dependem essencialmente da utilização dos dados, por oposição à utilização de conhecimento específico do domínio do problema, são também muitas vezes designados modelos *data-driven*. Este tipo de modelo tem conhecido grande sucesso, especialmente na resolução de problemas complexos, sendo caracterizados pela utilização de grandes conjuntos de dados. A premissa essencial dos métodos não-paramétricos é a de que relações que ocorrem de forma consistente no conjunto de dados repetir-se-ão em observações futuras; uma perspectiva eminentemente indutiva. Um importante benefício decorrente deste tipo de modelos é o de que não exigem um conhecimento profundo do fenómeno a modelar, facto particularmente útil no tratamento de problemas espaciais, dada a sua complexidade.

Neste contexto é indispensável introduzir o conceito de pré-processamento, que não é mais do que *a remoção de informação irrelevante e a extracção das características essenciais por forma a simplificar o problema* (Liu *et al.*, 1998). Com base neste conceito surgem os modelos não-paramétricos com pré-processamento, que podem ser encarados como o meio-caminho entre os modelos paramétricos e os modelos não-paramétricos. Estes podem ser vistos como a separação do desenvolvimento do modelo em duas partes distintas: aplicação de conhecimento específico do domínio de estudo (pré-processador); aspectos menos bem compreendidos do problema (modelo). Estes modelos possibilitam, numa primeira fase, a utilização de conhecimento prévio que possuímos sobre o fenómeno, por exemplo para avaliar a interacção entre duas cidades necessitamos de saber o número de habitantes de cada uma, bem como a distância que as separa (modelo gravitacional). Na segunda fase, o modelo é aplicado de maneira a especificar a forma como as variáveis se relacionam e a obter o *output* desejado. Existe, assim, por um lado a necessidade de seleccionar e adequar as variáveis relevantes e, por outro, a de especificar a forma como cada uma delas contribui para o resultado final.

III. *DATA MINING* GEO-ESPACIAL

Quando falamos da relação entre a Geografia e DM, um paradoxo pode facilmente ser observado. Por um lado, o DM é, neste momento, encarado, por grande parte da comunidade da Ciência da Informação Geográfica (CIG), como a grande esperança para aumentar a qualidade e sofisticação da análise espacial. Por outro lado, podemos argumentar que há muito que a Geografia pratica aquilo que hoje se denomina DM. De facto, se aceitarmos que a principal função do DM consiste em melhorar as interfaces entre

os sistemas de armazenagem de dados e os humanos, proporcionando a exploração, resumo e modelação de grandes bases de dados, então há séculos que a Geografia faz DM. Exemplo disso é o mapa de Charles Minard (fig.1).

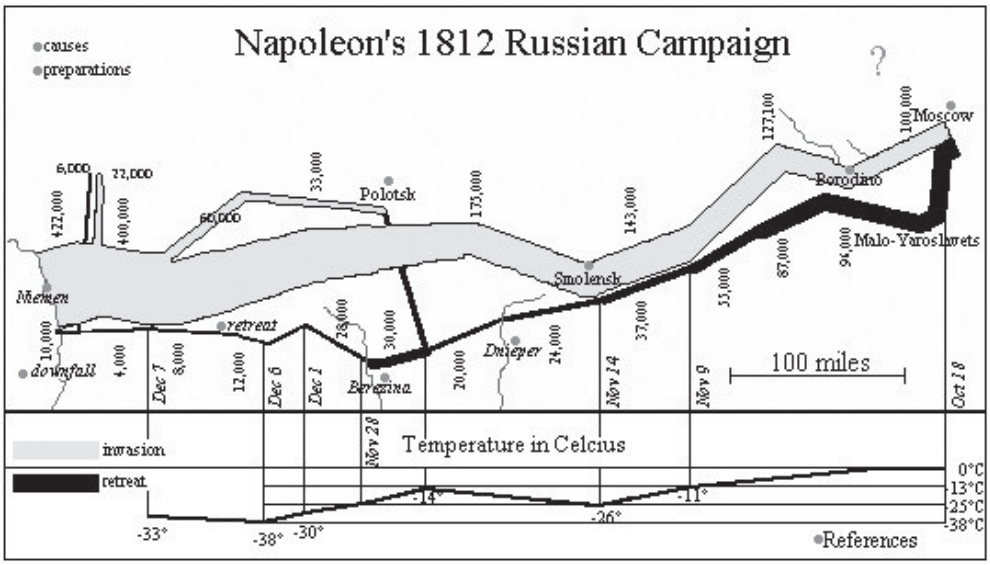

Fig. 1 – Campanha russa de Napoleão (adaptado de Burch e Grudnitski, 1989).

Fig. 1 – The Russian campaign of Napoleon (adapted from Burch e Grudnitski, 1989).

Esta obra prima da cartografia refere-se à campanha de Napoleão na Rússia, e pensamos que dá uma boa *imagem* das capacidades dos mapas para sintetizar informação e como interfaces entre os seres humanos e grandes bases de dados. Outro bom exemplo deste argumento foi o trabalho desenvolvido pelo Dr. John Snow, aquando da grande epidemia de cólera em 1854 em Londres. Esta é talvez a mais fantástica descoberta que os SIG produziram, paradoxalmente, muito antes da existência de computadores.

A segunda parte do paradoxo de que inicialmente falávamos relaciona-se com o facto de que há muito que o desequilíbrio entre as capacidades de armazenamento, gestão e acesso e as ferramentas de análise proporcionadas pelos SIG foi notado (Aangeenbrug, 1991; Openshaw, 1993; Anselin, 1993; Goodchild, 1991). De facto, a sofisticação oferecida pela evolução da Ciência Computacional, no âmbito do acesso e gestão dos dados geográficos, não tem paralelo na análise espacial. Por isso temos SIG com capacidade para armazenar e gerir grandes quantidades de dados geo-referenciados, mas não possuímos as ferramentas que permitam transformar estes dados em informação e esta informação em conhecimento.

Por um lado, o mapa, ou de forma mais genérica as ferramentas de visualização proporcionadas pelos SIG, constitui uma ferramenta central na área da visualização de fenómenos complexos, nomeadamente aqueles que se manifestam na superfície terrestre (Bracken, 1994; Brodlie, 1994; Dorling, 1994; Visvalingam, 1994). Por outro lado, parece indispensável que a Geografia dê atenção às novas ferramentas que vão surgindo no âmbito do DM e que podem contribuir de forma decisiva para uma nova era na Geogra-

Fig. 2 – Distribuição das vítimas de cólera e poços de recolha de água em Londres (Snow, 1854) (adaptado de Tufte, 1983).

Fig. 2 – The distribution of cholera victims and water pumps in London as drawn by Dr. John Snow in 1854 (adapted from Tufte, 1983).

fia. Esta nova era deverá ser caracterizada por uma maior capacidade de previsão da evolução dos fenómenos estudados.

O aspecto mais marcante da Geografia hoje em dia é a explosão de dados geo-referenciados (Openshaw, 2000b) produzida pelos recentes desenvolvimentos nas Tecnologias de Informação. Tecnologias de recolha de informação com referências geográficas, que vão desde a Cartografia Digital e Detecção Remota até aos *Location Based Services* (LBS), têm vindo a «inundar» as bases de dados. Este facto realça a importância do desenvolvimento de ferramentas capazes de lidar, de forma efectiva, com grandes quantidades de dados geo-referenciados. Hoje, necessitamos de ferramentas capazes de fazer face ao carácter multivariado e altamente complexo dos dados. A questão pode ser formulada dizendo que em termos de CIG vivemos num ambiente «rico em dados e pobre em teoria» (Openshaw, 1993). No cenário actual temos os dados para responder a muitas e urgentes questões (sociais e ambientais) mas pouco mais do que os tradicionais *buffers* e *overlays* para os analisar. O DM constitui para a Geografia uma oportunidade para levar a análise espacial a outros níveis de sofisticação, promovendo uma exploração mais efectiva das bases de dados disponíveis.

IV. ENQUADRAMENTO CIENTÍFICO DO *DATA MINING* GEO-ESPACIAL

Como já dissemos, as ferramentas que caracterizam o DM possuem um cariz fortemente indutivo, pouco adequado ao trabalho científico. Este é um dos primeiros desafios que se põe a quem pretenda utilizar o DM no âmbito da Geografia. É indispensável definir metodologias cientificamente consistentes que permitam a utilização do DM como ferramenta para a descoberta científica.

Decorrente do argumento apresentado na secção II sobre os modelos utilizados em ciência, podemos dizer que a abordagem determinística e a abordagem não-paramétrica à modelação constituem dois extremos de um contínuo. A abordagem determinística recorre apenas a conhecimento existente sobre o fenómeno; os modelos não-paramétricos recorrem a grandes quantidades de dados e capacidade de processamento, como forma de suprir a ausência de conhecimento. Os modelos paramétricos e não-paramétricos com pré-processamento podem ser vistos como situações de compromisso entre estes dois extremos, tal como se observa na figura 3.

Fig. 3 – O contínuo entre os modelos determinísticos e os modelos não-paramétricos.
Fig. 3 – The continuum between deterministic models and non-parametric models.

Pensamos que o sucesso passará sempre por conseguir, progressivamente, transportar os problemas científicos do lado direito da figura (abordagem não-paramétrica) para o lado esquerdo (abordagem determinística). Por outras palavras, o sucesso reside na capacidade de utilizar os dados e o poder computacional para produzir conhecimento passível de ser generalizável. Apesar de esta constituir a situação ideal, isto não significa que seja atingível em todas as ciências. Um exemplo, particularmente relevante, relaciona-se com a evolução da Física, para muitos o paradigma da ciência. Baseada essencialmente numa visão determinística dos fenómenos, tem vindo, progressivamente, em campos específicos, a adoptar abordagens probabilísticas. Os modelos determinísticos têm-se mostrado incapazes de acomodar, com as suas explicações, novos factos que derivam em parte da melhoria dos instrumentos de observação e medição.

A evolução da técnica e das ferramentas proporciona, muitas vezes, soluções que a ciência ainda não está preparada para oferecerda figura. Este talvez seja o actual estado da Geografia e da CIG com respeito ao DM. No século XVIII, o *Comité de Longitud*e, criado pelo parlamento britânico, foi obrigado a dar o prémio para a medição de longitude no mar a um relojoeiro em detrimento dos astrónomos do reino, que investiram fortemente na competição. Apesar dos constantes boicotes ao aparelho de Harrison, o comité, composto por astrónomos, não pôde contornar a evidência: o relógio de Harrison era muito mais preciso do que as tabelas de dados referentes à posição da Lua relativamente a

outros corpos celestes propostas pelos astrónomos (Macmillan, 1997; Sobel, 1995). Hoje a Geografia possui, através do DM, ferramentas que permitem a resolução de muitos problemas, para os quais ainda não existem respostas científicas.

Existe alguma unanimidade na comunidade da CIG de que, quer os modelos determinísticos, quer os modelos paramétricos, têm pouco a oferecer, nesta fase, para o desenvolvimento da Geografia e da CIG (Openshaw, 1993; Anselin, 1993). Este tipo de modelos exige um tipo de conhecimento e de formulações teóricas sobre fenómenos que pura e simplesmente ainda não existe; é indispensável construí-lo. A discussão deve centrar-se na forma como podemos utilizar os modelos não-paramétricos, para resolver problemas práticos da Geografia, mas também, para desenvolver as bases teóricas que sirvam ao desenvolvimento de uma nova Geografia. Existem diferentes perspectivas sobre a forma como a Geografia pode beneficiar da utilização de modelos não-paramétricos (Gahegan, 2001; Openshaw, 1999; Openshaw *et al*., 2000a).

A perspectiva defendida por Openshaw (2000a) centra-se na utilização do DM como uma «caixa preta», traduzindo-se na recolha de todas as variáveis explicativas disponíveis e deixando que as ferramentas processem toda a informação e apresentem uma solução final, não havendo acesso às especificações que nos permitam compreender o funcionamento do modelo (Openshaw, 2000b; Openshaw *et al*., 2000a). Esta perspectiva promove a análise automática da informação sem (ou com muito pouca) intervenção humana. Apesar de bastante pragmática e apta a proporcionar soluções para diversos problemas práticos, sofre de uma falha fundamental: é pouco apropriada ao trabalho científico, não proporcionando qualquer melhoria na compreensão da realidade. Esta visão indutiva é, em nossa opinião, insuficiente, tal como Popper refere (citado por Baudouin, 1989) *pouco importa o grande número de cisnes brancos que tenhamos observado; não justifica a conclusão de que todos os cisnes são brancos*. Um grande número de enunciados singulares nunca permite inferir um enunciado geral. Tal como Russel (citado em Deutsch, 1998) observa, uma galinha indutiva considera que o seu dono está genuinamente interessado no seu bem-estar, na medida em que todos os dias a alimenta, prevendo que o dono lhe continuará a trazer comida, ficando bastante surpreendida no dia em que o dono lhe corta o pescoço.

Tal como Popper a vê, a descoberta científica é governada por uma lógica invariável que inclui três momentos sucessivos. Num primeiro tempo, o cientista constrói cenários, conjecturas ou hipóteses, ou seja tentativas (*trial*) com vista a resolver os inúmeros problemas que a complexidade do universo lhe sugere. Num segundo tempo, submete as suas conjecturas a apertados testes que têm como objectivo refutá-las. Por fim, o método *trial and error* implica a renúncia, por parte do cientista, às certezas individuais e a aceitação de que as suas conjecturas sejam alvo de debate público e combatidas no seio da comunidade científica. De salientar a importância da refutação na teoria *popperiana* da descoberta científica; nenhuma teoria ou conjectura pode ser verificada, apenas ainda não foi refutada. Uma teoria será científica, de acordo com Popper, caso exista a possibilidade de a refutar ou testar. Neste contexto uma teoria nunca é mais do que uma hipótese que ainda não foi refutada, tendo por isso um carácter provisório até que novos factos a refutem ou novas teorias a substituam.

A nossa perspectiva sobre a utilização do DM tenta conciliar a abordagem *popperiana* com o processo do DM. Para isso é atribuído um papel decisivo à fase do pré-processamento, onde se procede à redução e eventual transformação do espaço de *input*, permitindo maior controlo sobre a especificação do problema, o que conduzirá a uma compreensão mais profunda do mesmo. Assim, as ferramentas de DM serão utilizadas no segundo passo enunciado na teoria de Popper, como forma de sujeitar as conjecturas aos

testes que eventualmente as refutem. É indispensável que haja um conjunto de hipóteses sobre o problema em apreço, estas devem proporcionar uma explicação compreensível e testável do fenómeno. O argumento científico nasce na conjectura do investigador e não por sugestão de um procedimento de busca automático, a conjectura é fundamental como actividade inicial.

No caso da utilização dos modelos não-paramétricos aceitamos passivamente o desconhecimento sobre o problema e esperamos que a tecnologia, por via da «força bruta» de busca no espaço das soluções, nos apresente a melhor solução. No caso da utilização dos modelos não-paramétricos com préprocessamento, é possível formular hipóteses, conjecturas quanto às variáveis que influenciam o fenómeno em estudo.

V. DISCUSSÃO E CONCLUSÃO

Em termos do desenvolvimento teórico, as duas abordagens expostas traduzem consequências completamente diferentes. Adoptando uma abordagem puramente indutiva não poderemos esperar aprender muito sobre os fenómenos que estudamos, apenas responder pontualmente a problemas práticos. Paralelamente, corremos o risco de chegar a relações espúrias, ou seja, relações que se verificam por acaso e não correspondem a nenhuma relação efectiva. De facto, à medida que aumentamos o número de variáveis independentes a probabilidade de ocorrência de relações espúrias aumenta. No segundo caso, pelo contrário, é possível antever o enriquecimento teórico, na medida em que, através da formulação de hipóteses sobre o comportamento dos fenómenos em apreço e o seu teste, poderemos compreender melhor quais as variáveis determinantes no processo e a forma como interagem. Esta abordagem experimental constitui um compromisso entre o conhecimento necessário para especificar um modelo paramétrico e a abordagem não-paramétrica.

Pensamos que só através de uma filosofia orientada para o fortalecimento teórico e conceptual poderemos antever contribuições significativas para a resolução do problema do *deficit* de ferramentas de análise espacial, que actualmente assola a Geografia e os SIG. Eventualmente, terá sido a ausência deste tipo de preocupações que levou ao insucesso da Revolução Quantitativa na Geografia Anglo-Saxónica. A adopção acrítica de métodos estatísticos (Anselin, 1989; Bailey, 1994; Gahegan, 2001; Openshaw, 1993), em grande parte desapropriados, resultou numa total incapacidade de generalização dos resultados produzidos.

É importante compreender que a *Descoberta de Conhecimento*, na sua versão mais automatizada, pode ser um excelente auxiliar do investigador, proporcionando interrogações conducentes a novas hipóteses e pistas para novas investigações. No entanto, esta é uma fase que não se encontra no âmbito do processo de desenvolvimento científico, sendo antes uma ferramenta auxiliar do cientista. O ressurgimento da noção de abdução (Gahegan, 2001), como o acto simultâneo de descobrir estruturas nos dados e produzir as hipóteses que a explicam, denota o esforço de enquadramento epistemológico desta nova área de conhecimento. No entanto, este é um trabalho em curso e que ainda se encontra longe de um quadro completo e coerente.

Pensamos ser necessária a formulação de hipóteses passíveis de ser testadas bem como formas inventivas de integrar referências geográficas, quer nos dados quer nas metodologias. A distância e as matrizes de conectividade são apenas dois exemplos das diversas formas que podemos utilizar para contextualizar os dados espaciais. Por exemplo, a análise de *clusters* há muito que é uma importante ferramenta do arsenal dos

geógrafos. Recentemente, tem sido muito utilizada naquilo que normalmente se designa a Geodemografia. A análise de *clusters* é particularmente interessante para a Geografia, na medida em que, ao contrário de outros métodos estatísticos, não assume qualquer hipótese sobre o tipo de distribuição dos dados. Neste sentido podemos considerá-la uma ferramenta bastante segura em termos da sua utilização com dados geo-referenciados. O problema encontra-se na forma como a maior parte dos geógrafos a utiliza. Seria de esperar que os geógrafos introduzissem algum tipo de medida ou indicador espacial na análise. Em vez de analisar a proximidade dos indivíduos em termos de um espaço construído pelas variáveis alfanuméricas, porque não introduzir a distância geográfica?

Por fim é importante salientar que a preocupação essencial do geógrafo deve ser a de salientar o espaço enquanto factor central das explicações obtidas. Assim, a localização, a área, a topologia, o arranjo espacial, a distância e a posição constituem o núcleo da pesquisa.

BIBLIOGRAFIA

Aangeenbrug, R. (1991) – A critique of GIS. *In* Maguire, D. J.; Goodchild, M. F. and Rhind, D. W. (eds.) – *Geographical Information Systems*, Vol. 1. Overview. Longman Scientific & Technical, Harlow: 101-107.

Abler, R.; Adams, J. S. and Gould, P. (1977) – *Spatial Organization, The Geographer's View of the World*. Prentice-Hall International, Inc., London.

Anselin, L. (1989) – *What is Special About Spatial Data? Alternative Perspectives on Spatial Data Analysis*, Technical Paper, NCGIA. Geography Departement, University of California Santa Barbara.

Anselin, L. (1993) – Exploratory spatial data analysis and geographic information systems. Proceedings of the workshop on *New Tools for Spatial Analysis*. ISEGI, Lisboa: 45-54.

Bação, F. and Painho, M. (2002) – *A Point Approach to Zone Design*. Proceedings from the 5th AGILE Conference on Geographic Information Science. Balearic Islands.

Bailey, T. (1994) – A review of statistical spatial analysis in geographical information systems. *In* Fotheringham, A. S. and Rogerson, P. A. (eds.) – *Spatial analysis and GIS*. Taylor and Francis Ltd., London: 13-44.

Baudouin, J. (1989) – *Karl Popper*. Biblioteca Básica de Filosofia. Edições 70, Lisboa.

Bird, J. (1993) – *The changing worlds of Geography, a critical guide to concepts and methods*. Second Edition. Clarendon Press, Oxford.

Bracken, I. (1994) – Towards improved visualization of socioeconomic data. *In* Hearnshaw, H. M. and Unwin, D. J. (eds.) – *Visualization In Geographical Information Systems*. John Wiley & Sons Ltd., New York: 76-84.

Brodlie, K. (1994) – A typology for scientific visualization. *In* Hearnshaw, H. M. and Unwin, D. J. (eds.) – *Visualization in Geographical Information Systems*. John Wiley & Sons Ltd., New York: 34-41.

Burch, J. and Grudnitski, G. (1989) – *Information Systems – Theory and Practice*. Fifth Edition. John Wiley & Sons, New York.

Couclelis, H. (1998) – Geocomputation in context. *In* Longley, P. A. *et al.* (eds.) – *Geocomputation: A Primer*. Wiley, Chichester: 17-30.

DEUTSCH, D. (1998) – *The Fabric of Reality*. Penguin, New York.

DORLING, D. (1994) – Cartograms for visualizing human geography. *In* HEARNSHAW, H. M. and UNWIN, D. J. (eds.) – *Visualization in Geographical Information Systems*. John Wiley & Sons Ltd., New York: 65-76.

FAYYAD, U.; SHAPIRO, G.; SMITH, P. and UTHURUSAMY, R. (1996) (eds.) – *Advances in Knowledge Discovery and Data Mining*. AAAI Press, Menlo Park, California.

GAHEGAN, M. (2001) – *Data mining and knowledge discovery in the geographical domain*, disponível no URL: http://www7.nationalacademies.org/cstb/wp_geo_gahegan.pdf (data da última consulta 15/09/2003).

GETIS, A. (1994) – Spatial dependence and heterogeneity and proximal databases. *In* FOTHERINGHAM, A. S. and ROGERSON, P. A. (eds.) – *Spatial analysis and GIS*. Taylor and Francis Ltd., London: 105-119.

GOODCHILD, M. F. (1991) – Guest Commentary: Geographic Information Systems. *Journal of Retailing*, 67: 3-15.

HAN, J. and KAMBER, M. (2001) – *Data Mining – Concepts and Techniques*. Morgan Kaufmann, San Francisco, California:

HAND, D. (1999) – Statistics and data mining: intersecting disciplines. *SIGKDD Explorations*, 1: 16-19.

JOHNSTON, R. J. (1986) – *Geografia e Geógrafos*. Difel, São Paulo.

KENNEDY, R.; LEE, Y.; ROY, B.; REED, C. and LIPPMANN, R. (1998) – *Solving Data Mining. Problems through Pattern Recognition*. Prentice Hall, Upper Saddle River, New Jersey.

LIU, H.; MOTODA, H. and YU, L. (1998) – Feature Exctraction, Selection, and Construction. *In* LIU, H. and MOTODA, H. (Ed.) – *Feature Extraction, Construction, and Selection: A Data Mining Perspective*. New York.

MACMILLAN, W. (1997) – Computing and the science of Geography: the postmodern turn and the geocomputational twist, *Geocomputation 1997*, disponível no URL: http://www.geocomputation.org/1997/papers/macmillan.pdf (data da última consulta 15/09/2003).

OPENSHAW, S. (1991) – Developing appropriate spatial analysis methods for GIS. *In* MAGUIRE, D. J.; GOODCHILD, M. F. and RHIND, D. W. (eds.) – *Geographical Information Systems*, Vol. 1 - *Principles*. Longman Scientific & Technical, Harlow: 389-402.

OPENSHAW, S. (1993) – *What is gisable spatial analysis?*. Proceedings of the workshop on *New Tools for Spatial Analysis*. ISEGI, Lisboa: 36-44.

OPENSHAW, S. (1994) – Two exploratory space-time-attribute pattern analysers relevant to GIS. *In* FOTHERINGHAM, A. S. and ROGERSON, P. A. (eds.) – *Spatial analysis and GIS*. Taylor and Francis Ltd., London: 83-103.

OPENSHAW, S. (1999) – Geographical data mining: key design issues, *GeoComputation 99*, disponível no URL: http://www.geovista.psu.edu/sites/geocomp99/Gc99/051/gc_051.htm (data da última consulta 15/09/2003).

OPENSHAW, S.; FISHER, M.; BENWELL, G. and MACMILLAN, B. (2000a) – GeoComputation research agendas and futures. *In* OPENSHAW, S. and ABRAHART, R. (eds.) – *GeoComputation*. Taylor and Francis Ltd., London: 379-401.

OPENSHAW, S. (2000b) – GeoComputation. *In* OPENSHAW S. and ABRAHART, R. (eds.) – *GeoComputation*. Taylor and Francis Ltd., London: 1-33.

O'KELLY, Morton E. (1994) – Spatial analysis and GIS. *In* FOTHERINGHAM, A. S. and ROGERSON, P. A. (eds.) – *Spatial analysis and GIS*. Taylor and Francis Ltd., London: 65-79.

ROGERSON, P. A. and FOTHERINGHAM, A. S. (1994) – GIS and spatial analysis: introduction and overview. *In* FOTHERINGHAM, A. S. and ROGERSON, P. A. (eds.) – *Spatial analysis and GIS*. Taylor and Francis Ltd., London: 1-10.

SOBEL, D. (1995) – *Longitude*. Fourth Estate, London.

SNOW, J. (1854) – *On the Mode of Communication of Cholera*, 2nd edition. John Churchill, New Burlington Street, London, England, 1855. Published by CHEFFINS, C. F. Lith, Southhampton Buildings, London, England. Disponível em http://www.ph.ucla.edu/epi/snow/snowmap1_1854_lge.htm (data da última consulta 15/09/2003).

TUFTE, E. R. (1983) – *The visual display of quantitative information*. Graphics Press, Cheshire.

VISVALINGAM, M. (1994) – Visualisation in GIS, Cartography and ViSC. *In*: HEARNSHAW, H. M. and UNWIN, D. J. (eds.) – *Visualization In Geographical Information Systems*. John Wiley & Sons Ltd., New York: 18-25.

WEISS, S.; INDURKHYA, N. (1998) – *Predictive Data Mining – A Pratical Guide*. Morgan Kaufmann, San Francisco, California.