Capítulo 5
MÉTODOS E TÉCNICAS DE ANÁLISE DE DADOS FREQUENTEMENTE APLICADOS EM GEOGRAFIA DA SAÚDE

5.1 - O significado da análise de dados geográficos

Como pudemos verificar em capítulos anteriores, quer ao nível da produção cartográfica, quer ao nível do desenvolvimento dos estudos ecológicos, foram sintetizadas algumas regras que do ponto de vista epidemiológico e geográfico são úteis e necessárias à validação das investigações desenvolvidas. Sublinhou-se também a relevância afirmativa que a aplicação da moderna metodologia cartográfica detém quando associada a ferramentas quantitativas de análise e exploração de dados, contribuindo para evidenciar o papel que o espaço e o arranjo espacial detém ao nível da compreensão dos resultados em saúde. Por este motivo, procede-se agora à apresentação de um conjunto de métodos e técnicas aplicadas ao nível da geografia na abordagem gráfica e quantitativa dos fenómenos de saúde, directa ou indirectamente relacionados com os processos de distribuição e difusão espacial.

Previamente, deve assinalar-se que os diferentes dados relacionados com a saúde podem ser submetidos a um conjunto de análises mais ou menos detalhadas que incluem, ou não, a dimensão espacial dos fenómenos, gerando importantes diferenças entre a análise espacial de dados de saúde e a análise de dados de saúde.

Por *análise*, pode entender-se, em sentido lato, uma avaliação inicial dos dados, bem como a aplicação de técnicas analíticas adequadas à interrogação dos mesmos (Robinson, 1998). Neste contexto, e tendo presente a valorização da componente espacial, sublinha-se a aplicabilidade de dois tipos de análise: a análise geográfica e a análise estatística espacial.

No caso da *análise geográfica* em saúde, assume particular relevo a investigação das relações topológicas de *proximidade* entre os dados (fontes de contaminação), *conectividade* (acesso a serviços de saúde, envolvendo estudos de acesso físico), *adjacência* (estudos de vizinhança) e *contingência* (quantidade de eventos por unidade de área). Para além desta, e encetando desejáveis relações de complementaridade ao nível da produção de informação em saúde, emerge a *análise estatística espacial* que,

partindo de dados ou de indicadores de risco frequentemente construídos em saúde, incorpora a informação sobre a localização espacial dos dados na formulação de modelos e técnicas de análise estatística onde a referência geográfica é deliberadamente e explicitamente utilizada (Bailey e Gatrell, 1995; Assunção 2001). Contrariamente, quando a informação espacial não é objecto de utilização em nenhum momento do processo de análise, independentemente da sofisticação das ferramentas estatísticas aplicadas, estamos no âmbito exclusivo da *análise de dados* o que, apesar de útil e necessária, não permite evidenciar um conjunto de dinâmicas que são inerentes ao arranjo espacial dos eventos investigados.

Neste âmbito, Bailey e Gatrell (1995), Assunção (2001) entre outros, ao proporem uma *abordagem quantitativa* de dados geográficos relacionados com a distribuição espacial dos fenómenos de saúde, particularmente ao nível da elaboração cartográfica (*mapping disease*), assinalam a necessidade de se percorrerem três etapas fundamentais:

1. *Visualização*: compreende a elaboração de um sistema de dados quantitativos espacialmente referenciados, onde o processo de construção cartográfico assume indiscutível utilidade, permitindo evidenciar (ir)regularidades espaciais dos fenómenos estudados, para além de potenciar a sinalização de eventuais associações entre as aspectos sócio-ambientais adjacentes às variáveis investigadas. É certamente por este motivo que Meade (2000; pp. 10) sublinha a singularidade e a relevância das ferramentas cartográficas aplicadas aos estudos de morbilidade citando neste contexto um velho adágio: «Se não se pode cartografar não é geográfico…»;
2. *Análise exploratória de dados* (AED): segundo os autores citados, este patamar envolve a aplicação de um conjunto de procedimentos simples de descrição numérica, tais como medidas de dispersão e de centralidade em conjugação com imaginativas representações gráficas. A inclusão de procedimentos geográficos e de estatística espacial aplicada à AED, com o objectivo de efectuar uma prospecção sistemática e rigorosa de eventuais padrões espaciais, bem como detectar e verificar associações espaciais entre as diferentes variáveis que caracterizam os fenómenos estudados definem o que se designa por *análise exploratória de dados espaciais* (AEDE);
3. *Modelação*: Como estádio final, o processo de modelação envolve um conjunto diferenciado de ferramentas estatísticas, metodologicamente adaptadas às

diferentes características dos dados em análise, envolvendo, na sua fase final, o teste de hipóteses.

Relativamente ao primeiro estádio, *visualização*, independentemente da tipologia dos dados espacialmente referenciados – dados pontuais, linhas ou áreas, facto que adiante detalharemos, importa reconhecer que a maioria dos autores (Bailey e Gatrell, 1995; Robinson, 1998; Gatrell, 2002) assinalam a existência de uma grande fluidez e quase interdependência de procedimentos existente entre os diferentes níveis traçados, particularmente entre os procedimentos de *visualização* e *análise exploratória de dados*, sendo que, no decurso da apresentação gráfica dos dados quantitativos, o processo descritivo exige, quase em simultâneo, a manipulação de algumas variáveis, tais como intervalos de classe ou escala de análise, prefigurando a emergência de hipóteses que, posteriormente, devem ser objecto de análise e validação. Por este motivo, Robinson (1998) não procede à distinção entre o processo de visualização e análise exploratória de dados porquanto que, a elevada interactividade de procedimentos transforma-os na «outra face da mesma moeda». Para o autor, o processo de *análise exploratória de dados* que a geografia deve agora praticar, está algo distanciado do procedimento teórico-analítico da revolução quantitativa, quase exclusivamente focada em processos de análise confirmatórios, centrados na validação de pressupostos apriorísticos. Catell (1996; citado por Robinson, 1998; pp. 6) definiu este processo como «espiral indutiva-hipotético-dedutiva», onde os investigadores apenas questionavam vagamente as relações e as diferenças entre as variáveis de diferentes lugares, num processo de «*empirismo especulativo*», mais centrado na sofisticação de procedimentos analíticos do que na confrontação dos resultados com o real.

É pois nesta perspectiva que a *análise exploratória de dados* deve assumir particular significado, contribuindo para uma investigação geográfica mais clarividente, na medida que permite uma abordagem holística e coerente entre o fenómeno e as demais variáveis que lhe estão adjacentes e que, de um modo mais ou menos intenso, acabam por o modelar. Para Robinson (1998) a *análise exploratória de dados* enfatiza a importância de se examinarem os dados disponíveis de uma forma cuidada, proporcionando ao investigador um íntimo e importante relacionamento com os dados antes de se prosseguir adiante na investigação:

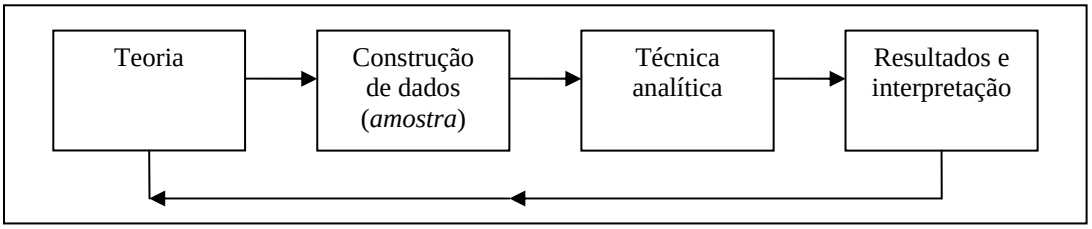
> «É uma ajuda para que se evitem muitos dos erros cometidos no passado por aqueles que se envolviam em precipitadas e complexas análises de dados espaciais, ignorando muitos dos problemas associados a este tipo de abordagem.» (Robinson, 1998; pp. 14).

Prosseguindo na relevância concedida a este patamar vital da investigação, o autor reforça a sua perspectiva recorrendo a uma poderosa metáfora utilizada por K. Jones (1984) neste contexto:

> «A exploração de dados funciona como um detective numérico, agrupando e peneirando através dos dados, como se estivesse a produzir prova para o procedimento judicial da análise confirmatória.»

Assim, o procedimento da *análise exploratória de dados* é, simultaneamente, um processo de reflexão e de (re)construção de associações e hipóteses acessórias aos dados, na medida em que suscita uma atitude crítica em relação aos pressupostos teóricos de suporte, evitando inferências abusivas através de um processo de retroacção suportado pela análise confirmatória que Robinson articula no diagrama exposto (Fig. 5.1).

Figura 5.1 – Ciclo de investigação analítica



Fonte: Robinson, 1998; pp. 13

Idêntica posição é defendida por Johnston *et al.* (2000;pp. 249)[1], onde o autor releva o fortalecimento teórico do trabalho de investigação geográfica através da análise exploratória de dados, fundamentando-se em dois pontos principais:

1. Muita da investigação geográfica tem uma relativa fragilidade teórica e as expectativas empíricas são consequentemente imprecisas nos seus objectivos;

[1] Ver- *Exploratory data analysis*. *In* - The Dictionary of Human Geography (2000), Blackwell Ed.

2. Alguns dados utilizados por geógrafos são especificamente colhidos respeitando condições e regras experimentais, pelo que o investigador pode ter uma visão incompleta da sua estrutura.

Por último, o processo de *modelização* apresentado por Bailey e Gatrell (1995), envolve o que Robinson (1998) e Johnston (2000) denominam por análise confirmatória, uma vez que envolve o teste de hipóteses, sendo que, em Geografia da Saúde, muitos dos modelos aplicados implicam o uso explícito da variável "localização espacial", ou consideram indicadores geo-sócio-demográficos que distinguem uma área de outra (Gatrell, 2002). Uma vez mais, somos confrontados com a dificuldade já mencionada de estabelecer linhas divisórias entre os diferentes patamares de procedimentos, tanto mais que a interactividade tem vindo a ser operacionalmente reforçada pela incorporação de *software* específico, capaz de integrar ferramentas aplicáveis à totalidade dos processos mencionados: visualização, exploração e modelação de dados.

Neste ponto, não poderíamos deixar de mencionar os recentes avanços na cartografia, que reclamam progressivas actualizações dos processos computorizados de georeferenciamento, nomeadamente os que estão associados aos *Sistemas de Informação Geográfica* (SIG), que se afirmam como poderosa ferramenta na elaboração e análise de mapas temáticos, quantitativos e qualitativos, demonstrativos da variabilidade espacial dos fenómenos tratados, autorizando a interligação de eventos de origem diversa, espacialmente referenciados, integrando, de modo relativamente "amigável", factores sociais e ambientais, subsidiários dos fenómenos a investigar, características que adiante desenvolveremos.

5.2 - **Análise Exploratória de Dados Espaciais**

A consideração de métodos e técnicas de análise, incluindo procedimentos de expressão gráfica, constituem uma ferramenta indispensável ao processo de análise de dados em ciências sociais e, de um modo particular na Geografia da Saúde.

Na investigação de processos relacionados com a saúde encontramos elevada preponderância de dados associados a áreas ou, em alguns casos, dados associados a

pontos, cujo processo de visualização mais comum assenta na produção de mapas temáticos representando a variação espacial de uma determinada variável.

No caso dos *dados pontuais* espacialmente referenciados, podemos analisar os chamados *pontos verdadeiros* - aqueles que efectivamente detêm localizações exactas de acontecimentos (acidentes, fontes de contaminação ambiental), ou ainda analisar *pontos conceptuais*. Nestes casos, devemos ter presente que os dados referentes a uma área representam uma agregação de valores no interior dos limites traçados, não estando adstritos a nenhum ponto específico desse espaço, sendo por vezes útil e necessário recorrer a um "artifício de análise" que referencia esses mesmos dados a um ponto representativo da área do polígono em causa (ponto conceptual), por exemplo o centro geométrico, também chamado de centróide, com o objectivo de poder utilizar técnicas de visualização e análise de dados pontuais (Slocum, 1999). Assim, o processo de investigação clássico procura averiguar a identificação espacial de aglomerados de pontos (eventos) *versus* a sua distribuição aleatória, permitindo sucessivas aproximações à fonte etiológica suspeita ou já confirmada. Segundo Rodrigues (1993, Cap.III), neste tipo de estudos, onde se objectiva ilustrar determinada hipótese etiológica, como a existência de uma fonte poluidora ou emissora de radiações, é indispensável a utilização de unidades de análise homogéneas e de pequena dimensão, quer do ponto de vista geográfico quer no que respeita aos níveis de exposição. Pode ainda ser necessário recorrer à aplicação de técnicas de mapeamento específicas, como a interpolação de valores para gerar superfícies de densidade, ou ainda recorrer à utilização de áreas de influência (*buffers*), englobando níveis de dose-efeito semelhantes, entre outras (Rodrigues, 1993; Santos *et al.*, 2000; Gatrell, 2002).

A este nível, e para além do exemplo clássico da detecção de casos de cólera na área Londrina no século XIX, Bailey e Gatrell (1995) fornecem-nos o exemplo da difusão da infecção por *legionella* em Filadélfia (1978), onde a identificação de *clusters*[2] espaciais de incidência contribuiu, de forma determinante, para isolar o ponto suspeito de difusão da infecção.

Relativamente aos dados investigados por *unidade de área*, encontramos um determinado número de casos y_i adstritos a uma área geograficamente definida *i*, maioritariamente representados por (sub)divisões administrativas cujos limites são

[2] *Cluster* – a expressão pode traduzir um conjunto de casos, grupos ou eventos que parecem relacionar-se pela forma de distribuição no espaço ou no tempo, em quantidades maiores do que poderia ser esperado pelo acaso. A detecção destes agrupamentos espaciais é fundamentada por desvios estatísticos significativos medidos pelo valor de *p* (Last, 1988).

traçados em função de factores homogéneos (ex.: NUTS), ou arbitrariamente orientados por mera conveniência política.

Para além da tipologia dos dados, importa também considerar a aplicação de um conjunto de ferramentas que permitem descrever o comportamento das variáveis quantitativas, tornando mais inteligível e comparável o que graficamente se resume num mapa, seja ele de pontos, símbolos geométricos, mapa coropleto, mapa de fluxos, mapa de isolinhas, entre outros.

No caso da distribuição dos dados de saúde, recorre-se, maioritariamente à utilização de taxas e razões para expressar medidas de risco, posteriormente materializadas em mapas temáticos. Neste contexto, Assunção (2001) hierarquiza um conjunto de objectivos que devem orientar a produção e análise deste tipo de dados:

> 1 – *Descrição*: consiste no processo de visualização da distribuição espacial da doença em determinadas áreas;
>
> 2 – *Exploração*: sugere e averigua a existência de determinantes locais da patologia e respectivos factores etiológicos capazes de suportarem hipóteses posteriormente investigadas;
>
> 3 – *Investigação de associações*: averigua a significância estatística e o comportamento de factores etiológicos e/ou outros elementos relevantes presentes na área, capazes de potenciarem o risco, podendo ou não constituir factores causais.

Para se alcançar este conjunto de objectivos, diversos autores (Robinson, 1998; Assunção, 2001; Gatrell, 2002) mencionam a imprescindibilidade da utilização de ferramentas estatísticas descritivas e posteriormente analíticas, capazes de explicitarem e sustentarem adequadamente as hipóteses a formular, relevando o valor da análise quantitativa em três pontos fundamentais: auxiliar a consolidação de conclusões retiradas de avaliações qualitativas; facilitar a comparação de diferentes populações no espaço e no tempo, através da aplicação de critérios objectivos e reprodutíveis do ponto de vista quantitativo; permitir a emergência de características, ou conjuntos de características, que escapam a avaliações exclusivamente qualitativas (Assunção, 2001;pp.2).

Walford (1995) e Robinson (1998) sugerem que o processo de investigação quantitativa, ao nível da análise exploratória de dados, deve ser iniciado por um

conjunto de procedimentos simples, tais como o cálculo de medidas de centralidade e dispersão: moda, média e desvio-padrão, para além dos desvios à mediana expressos por intervalos interquartis.

Embora as medidas de centralidade sejam úteis para se conhecer uma distribuição, por si só são insuficientes, podendo adicionar-se o cálculo da co-variância e do coeficiente de correlação de modo a suportarem comparações explicativas entre duas variáveis (Ferreira e Simões, 1987).

A tipologia dos dados anteriormente apontada, só permite a aplicação de ferramentas mais sofisticadas, no âmbito da análise estatística espacial, quando estes estão referenciados por um sistema de coordenadas. Tal como exemplifica Assunção (2001, pp.3-5), a análise de uma regressão linear simples, utilizando como variável dependente a taxa de homicídio de uma dada área *versus* o PIB dessa mesma área, como variável independente, não constitui por si só uma análise estatística espacial, ainda que possamos verificar que os dados utilizados estão espacialmente referenciados.

Por outro lado, tendo presente a localização das áreas geográficas em análise e o seu arranjo espacial – adjacência, conectividade e contingência, não é muito razoável admitir que os valores das variáveis observadas em unidades contíguas sejam absolutamente independentes. Antes, deve esperar-se uma certa similaridade ou correlação entre unidades que detenham algum grau de vizinhança espacial. Tal como esclarecem Bailey e Gatrell (1995), Assunção (2001) e Câmara *et al.* (no prelo), a consideração da evidente continuidade das actividades humanas no espaço, apesar da arbitrariedade dos limites administrativos, autoriza a presunção apriorística daquilo que Waldo Tobler (1979), designou por primeira lei da geografia: «*todas as coisas são parecidas, mas coisas mais próximas parecem-se mais do que coisas mais distantes*». Tal como sublinha Assunção (2001, pp. 4), partindo desta premissa de similaridade de atributos entre áreas de proximidade, podemos melhorar as estimativas de parâmetros específicos de cada área recorrendo à redundância informativa dos vizinhos, i.e., a totalidade das observações em estudo estão envolvidas na descrição e averiguação de padrões ou tendências inerentes ao fenómeno estudado, sendo que nenhum dado – pontos, áreas ou linhas, deve ser interpretado como atributo independente dos demais. Por este motivo, e ao nível da análise exploratória de dados espaciais, a aplicação de ferramentas estatísticas que *incluem especificamente a variável "localização espacial" no seu algoritmo*, autoriza a detecção de padrões espaciais estatisticamente

significativos para os parâmetros considerados, ou permite-nos tão só detectar e sinalizar “regiões” cujos indicadores exibem valores semelhantes.

Como primeira etapa da AEDE, procurando averiguar o nível de dispersão dos dados espaciais em relação à mediana, bem como a verificação de valores máximos e mínimos alcançados pela distribuição, podemos recorrer à execução de um gráfico *boxplot* que, como o nome indica, apresenta graficamente os dados agrupados em caixas, que representam os *quartis*, evidenciando também a presença de valores atípicos (*outliers*), superiores e inferiores da distribuição, sabendo que 50% das ocorrências centrais estão localizadas entre o quartil superior e o quartil inferior, denominado por intervalo interquartil (IIQ). Os valores considerados como *outliers* são, normalmente, aqueles que excedem 1,5 vezes o IIQ (tanto acima do 3.º quartil, como abaixo do 1.º quartil) podendo, no entanto, atingir valores até 2,5 vezes o IIQ, dependendo do objectivo de estudo.

Como ferramenta de visualização associada, encontramos o *boxmap* ou seja, o mapeamento dos valores dos quartis registados nas respectivas áreas, normalmente associados a cores, assinalando também os *outliers* superiores e inferiores, o que confere grande significado e facilidade de observação à análise exploratória espacial.

Para além disto, Rodrigues (1993), a propósito da utilização de dados geográficos de mortalidade por tumores malignos em Portugal, alerta-nos para a importância da detecção, numa primeira fase, de *gradientes geográficos* e, posteriormente, para a detecção de fenómenos de *autocorrelação espacial*. A observação de *gradientes geográficos* compreende a observação de uma variação orientada e regular da repartição espacial de um fenómeno (Rodrigues, 1993; pp.74). No caso de um estudo inglês citado pelo autor, verificou-se a detecção de alguns gradientes geográficos para a patologia cardiovascular, acompanhados, na mesma orientação, por diversas características socioeconómicas e hábitos alimentares facilitadores deste tipo de eventos.

Somos pois colocados perante a necessidade de conhecer e considerar medida(s) de autocorrelação espacial. Neste ponto abordaremos apenas a autocorrelação espacial univariada, i.e. aquela onde se leva em conta uma única variável medida nas diferentes unidades geográficas investigadas, cujo prefixo “*auto*” indica que a medida de correlação é realizada com a mesma variável aleatória, medida em distintos locais do espaço.

A utilização deste instrumento permitir-nos-á averiguar a possível existência de um conjunto, ou conjuntos, de unidades geográficas vizinhas que evidenciam valores próximos, sinalizando a possível existência de um factor espacial estatisticamente significativo *versus* um carácter aleatório dos valores irregulares. Neste ponto Rodrigues (1993, pp. 76) adverte para a possibilidade da detecção de autocorrelação espacial significativa, não só baseada em valores de risco idênticos, mas também decorrente do facto de que estimativas mais precisas estão associadas a áreas mais populosas, admitindo como razoável que áreas contíguas tenham efectivos populacionais semelhantes, estabilizando por esta via os valores das medições efectuadas.

Deve também esclarecer-se que, consoante o objecto em causa, o nível de significância estatística do índice adoptado pode ser calculado de *forma global*, i.e. tendo em conta o valor da variável obtido em todas as unidades investigadas, ou pode ser considerado de *forma local*, avaliando a significância estatística do valor calculado para uma área a partir dos seus vizinhos, podendo ainda considerar-se diversos graus de vizinhança (ex.: vizinho de primeira ordem, vizinho de segunda ordem ou vizinhos num raio determinado).

Acrescenta-se que, regra geral, estes processos de quantificação assumem como naturalmente necessária a elaboração de uma matriz de vizinhança *w*, com n^2 elementos (matriz de proximidade espacial), adstrita a um sistema de ponderação onde cada elemento da matriz w_{ij} representa uma medida de proximidade espacial da área A_i em relação à área A_j (alguns programas de tratamento de dados espaciais processam de modo automático a matriz de proximidade espacial). Tal como sublinham a maioria dos autores (Bailey e Gatrell, 1995; Assunção, 2001), a matriz é considerada uma ferramenta essencial para caracterizar a distribuição espacial dos objectos onde a metodologia mais simples de escolha dos valores de w_{ij} é aquela que atribui a áreas contíguas o valor 1 (objectos com fronteira comum $w_{ij} = 1$) e valor 0 a áreas que não exibem qualquer contiguidade (objectos sem fronteira comum $w_{ij} = 0$).

Entre as medidas de autocorrelação espacial univariadas comummente referenciadas destaca-se o *Índice Global de Moran*, que data de 1950 (Assunção, 2001), também designado por Índice Global de Associação Espacial, que nos permite averiguar a existência de *dependência espacial* de uma dada variável em diferentes lugares ou, caso contrário, indica-nos uma distribuição espacialmente aleatória. Para a determinação da

autocorrelação espacial usando o *Índice Global de Moran* (ou *I de Moran*) entre vizinhos de primeira ordem, usa-se a seguinte fórmula:

$$I = \frac{n\sum w_{ij}\left(z_i - \bar{z}\right)\left(z_j - \bar{z}\right)}{S_0 \sum_i \left(z_i - \bar{z}\right)^2}$$

onde:

$$S_0 = \sum i \neq j \quad w_{ij}$$

n = n.º de áreas;
z_i = valor da variável considerada na área *i;*
$\bar{z}$ = valor médio da variável na região de estudo;
w_{ij} = elementos da matriz padronizada de proximidade espacial;
z_j = valor da variável considerada na área *j*;

(Assunção, 2001; pp. 26)

Através deste indicador, podemos averiguar o grau de autocorrelação espacial de uma variável z em diferentes áreas *i* e *j*, (z_i, z_j), ponderada pela proximidade geográfica medida por w_{ij}, onde o numerador evidencia a média dos produtos dos desvios das áreas *i* e *j* em relação à média global, e o denominador é uma medida de variabilidade dos desvios. Os valores de *I de Moran* correspondem ao declive da recta de regressão e, à semelhança de um coeficiente de correlação linear, normalmente variam entre 1 e -1, ou seja:

- O *I* é *positivo* quando existe *dependência espacial*, com os valores das áreas vizinhas a evidenciarem similaridade entre si. Quanto mais perto de 1, maior é a autocorrelação espacial, sendo o valor 1 atribuído a uma *autocorrelação positiva* (directa) perfeita;
- O *I* é *negativo* quando existe dependência espacial, mas os *valores das áreas vizinhas são dissemelhantes*. Quanto mais perto de -1, maior é a autocorrelação espacial, sendo o valor -1 atribuído a uma *autocorrelação negativa* (inversa) perfeita (este é um fenómeno raro de observar);
- No caso de *I* ser 0 ou próximo de 0, indica que as variáveis são *espacialmente independentes*, ou seja, *não existe autocorrelação* (Assunção, 2001).

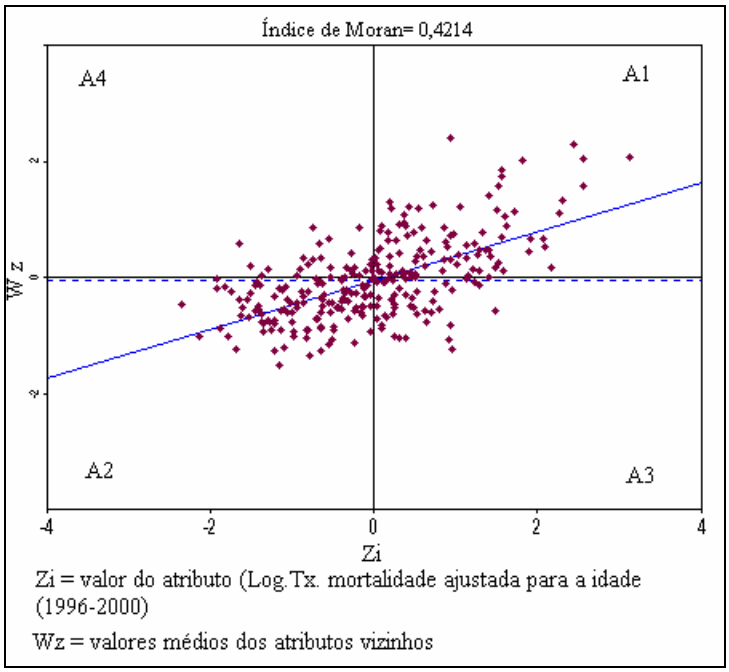
Tal como refere Câmara *et al.* [*prelo*], subjacente ao cálculo do índice está a consideração de duas hipóteses:

- *Hipótese nula* H_0: distribuição espacial aleatória dos dados obtidos para a variável em análise, ou seja, *independência espacial*;
- *Hipótese alternativa* H_1: existência de *dependência espacial*.

Neste contexto, é aconselhável estimar a validade estatística da autocorrelação espacial através da averiguação da significância do índice, sendo frequente associar a estatística do teste a uma distribuição normal. Se o "risco" admitido for por exemplo de 5%, significa que é de 0.05 a probabilidade (*p*) admitida de cometermos um erro associado à rejeição da hipótese nula.

Uma outra forma mais impressiva de observar dependência espacial pode ser conseguida através da construção do *Gráfico de Dispersão de Moran*, construído com base nos valores de uma distribuição normal de um atributo (z_i), comparados com a média dos valores dos atributos dos seus vizinhos w_{zi} (Neves *et al.*, s/d). Câmara *et al.* [*prelo*] fornece algum auxílio para a interpretação deste objecto, observando-o como um gráfico bidimensional para visualização da associação espacial entre *z* e *Wz* (média dos vizinhos), que se encontra dividido por quadrantes, tal como demonstra a figura 5.2.1, para a Taxa de Mortalidade por SIDA, Ajustada para a Idade para Portugal (continental), onde cada quadrante corresponde a um tipo de associação espacial, dependendo de como os valores *z* se relacionam com os vizinhos W_z e o valor do *I de Moran* corresponde à declividade da recta de regressão.

Figura 5.2.1 – Gráfico de Dispersão de Moran para o *log*. da Taxa de Mortalidade por Sida Ajustada para a Idade, Portugal continental (1996-2000)



- **A_1** – valores altos com médias dos vizinhos altas (*alto-alto*);
- **A_2** - valores baixos com médias dos vizinhos baixas (*baixo-baixo*); ou seja encontramos nos dois quadrantes, A_1 e A_2, pontos de associação espacial positiva ou directa, ou ainda *clusters* espaciais, já que determinada localização detém vizinhos com *valores idênticos* (Anselin, 1995);
- **A_3** – valores altos com médias dos vizinhos baixas (*alto-baixo*);
- **A_4** – valores baixos com médias dos vizinhos altas (*baixo-alto*); ou seja, evidencia associações espaciais negativas ou inversas, onde uma localização possui vizinhos com *valores dissemelhantes* dos seus – *outliers* espaciais (Anselin, 1995).

O que até aqui foi descrito, envolve uma medida global de autocorrelação (*Índice de Moran*) e o respectivo processo de visualização (*Gráfico de Dispersão de Moran*) o que, embora constitua um indicador útil, pode revelar alguma insuficiência, especialmente quando se investiga um elevado número de áreas associadas, como por exemplo a totalidade dos concelhos de uma NUT II ou III. Uma vez que o indicador agora considerado apenas fornece um único valor como medida de associação, alguns autores (Anselin, 1995) sugerem, com vantagem, a aplicação de *indicadores locais de autocorrelação*, como uma ferramenta estatística mais adequada para a detecção de sub-regiões com dependência espacial, permitindo identificar áreas de dependência espacial significativa (*hotspots*), que estão para além do alcance dos indicadores globais. Simultaneamente, não só se potencia a detecção de eventuais *clusters* espaciais, como também se autoriza a averiguação de uma ou mais relações de dependência espacial entre áreas dissemelhantes.

Uma das ferramentas de identificação de *autocorrelação espacial local* vulgarmente utilizada é o denominado *Índice Local de Moran* ou *Índice Local de Associação Espacial* – LISA (*Local Index of Spatial Association*), que permite determinar um índice de autocorrelação espacial para cada área, permitindo a identificação de agrupamentos, ou conjuntos de agrupamentos (*clusters*), onde a associação espacial é significativa, partindo da seguinte fórmula:

$$I_i = \frac{z_i \sum_{j=1}^{n} w_{ij} z_j}{\sum_{j=1}^{n} z_j^2}$$

n = n.º de áreas;
z_i = valor da variável considerada na área *i;*
$\bar{z}$ = valor médio da variável na região de estudo;
w_{ij} = elementos da matriz padronizada de proximidade espacial;
z_j = valor da variável considerada na área *j*;

(Câmara *et al.*, [*prelo*])

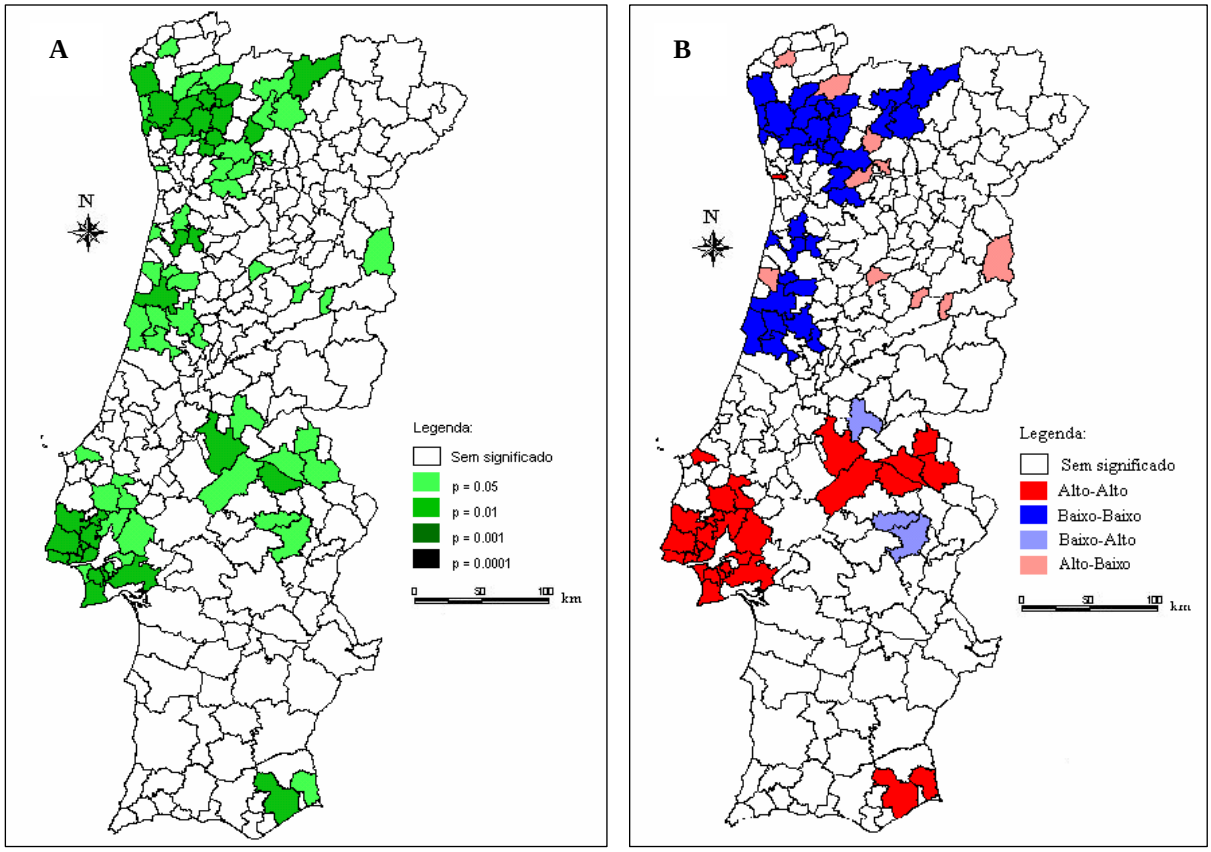
De acordo com o anteriormente descrito para o Índice Global de Moran, também a validade estatística do valor calculado para o *índice local* deve ser averiguada, verificando-se os respectivos valores de significância em relação à hipótese nula (independência espacial), sendo considerados estatisticamente significativos os valores < 0,05.

A ferramenta de visualização aplicada ao índice Moran local é conhecida por Mapa de Significância LISA (Fig. 5.2.2 A), onde são mapeadas e destacadas as *associações significantes* (p < 0,05). Normalmente, neste tipo de mapas as áreas são classificadas em quatro grupos: não significantes; com significância entre 0,05 e 0,01; com significância entre 0,01 e 0,001; e maior do que 0,001 (Anselin, 1995).

Existem ainda outras ferramentas de visualização como o Mapa de *clusters* LISA (Fig. 5.2.2 B) que classifica as áreas com autocorrelação espacial estatisticamente significativa, segundo os quadrantes do gráfico de dispersão de Moran a que pertencem.

Figura 5.2.2 - Mapa de *Significância* de Moran e Mapa de *clusters* Lisa:

A – Mapa de *Significância* Lisa baseado na significância do valor de *p* para o logaritmo da Taxa de Mortalidade por Sida Ajustada para a Idade – Portugal continental (1996-2000)
B - Mapa de *clusters* Lisa baseado na distribuição das áreas pelos respectivos quadrantes do gráfico de dispersão de Moran para o logaritmo da Taxa de Mortalidade por Sida Ajustada para a Idade – Portugal continental (1996-2000)



5.3 – Utilização de taxas específicas e taxas ajustadas de morbilidade e mortalidade

Quando se trabalham dados em saúde, seja em epidemiologia, geografia da saúde, ou em qualquer outra área do conhecimento, a forma de quantificar a morbilidade e a mortalidade assume vital importância ao nível da compreensão e interpretação do fenómeno estudado. Por esse motivo, os números reunidos sobre óbitos, casos de doença, acidentes, entre outros, são objecto de atenta observação, averiguando-se o seu impacto e o modo como se distribuem sob a população em risco, entendendo-se que

esta é constituída pela totalidade dos indivíduos que estão expostos aos respectivos eventos de morbilidade e mortalidade. Os eventos observados numa determinada população são, regra geral, expressos sob a forma de uma taxa, e constituem o numerador, sendo a população em risco colocada como denominador. A medida alcançada traduz o risco de doença ou óbito numa determinada população, para um intervalo de tempo definido, habitualmente um ano, sendo o seu valor multiplicado por uma constante (*K*), o que permite a sua indexação a um determinado número de pessoas (ex. 1.000 hab.; 10.000hab.; 100.000 hab.), (Mausner e Bahn, 1990; Stone *et al.*, 1999).

Entre as diversas medidas referenciadas para aferir a morbilidade numa dada população, duas delas são amplamente conhecidas: *taxa de incidência* e *taxa de prevalência*.

A *taxa de incidência* relaciona o número de novos casos de doença ocorridos numa determinada população:

$$Ti = \frac{y}{p} \times k$$

onde:

y = n.º novos casos de uma doença por período de tempo;
p = população do meio do período de tempo;
k = constante.

A *taxa de prevalência* mede o número de pessoas que num determinado intervalo de tempo são portadoras de determinada doença (incluindo casos já diagnosticados e novos casos):

$$Tp = \frac{c}{p} \times k$$

onde:

c= n.º de casos correntes de uma doença por período de tempo;
p = população do meio do período de tempo;
k = constante.

(Stone *et al.*, 1999; pp. 70)

A incidência reflecte, por esta via, a taxa de novos casos de doença que ocorrem num intervalo de tempo, em indivíduos expostos ao risco, traduzindo-se, simultaneamente, numa medida directa do risco de doença e da velocidade com que esta atinge uma população saudável, enquanto que a prevalência explicita o «número de casos de uma enfermidade que sobrecarregam uma população» num dado intervalo de tempo[3] (Stone *et al.*, 1999).

[3] Os indicadores apresentados, incidência e prevalência, também podem ser encontrados sob a forma de *razão*, uma vez que o termo *taxa* considera acontecimentos durante um período de tempo definido, não

Relativamente aos acontecimentos que envolvem óbitos, as *taxas brutas de mortalidade* expressam a relação existente entre a totalidade dos óbitos e a população em que os mesmos ocorrem, diferindo das que consideram no numerador causas específicas de obituário, originando *taxas de mortalidade por causa de morte*.
Outro tipo de taxas, com maior detalhe, compreende as taxas de mortalidade específicas para a idade, ou para outras componentes demográficas - raça e sexo, que explicitam no algoritmo a variável sobre a qual incide a taxa.

Taxa de mortalidade por causa de morte (Tm_c)

$$Tm_c = \frac{o_c}{p} \times k$$

onde:

o_c = n.º óbitos por causa específica por intervalo de tempo;

p = população do meio do período de tempo;

k = constante.

Taxa de mortalidade específica por idade (Tm_x)

$$Tm_x = \frac{o_x}{p_x} \times k$$

onde:

o_x = n.º óbitos por grupo etário por intervalo de tempo;

p_x = n.º indivíduos por grupo etário no meio do período de tempo;

k = constante.

Apesar da elevada utilidade dos indicadores apresentados, não podemos deixar de reflectir que as diferentes características da população em termos de estrutura etária, sexo e raça, comprometem a comparabilidade dos indicadores entre populações com composições diversas. Por exemplo, valores iguais ou semelhantes para a taxa de mortalidade por acidente vascular cerebral, mais frequente acima dos cinquenta anos, podem não traduzir o mesmo risco de morte para duas populações distintas, caso estas apresentem diferenças significativas ao nível da estrutura etária, nomeadamente quando

---

aferindo o processo de modo instantâneo. Neste caso referimo-nos a prevalência no momento e proporção de incidência.
Prevalência no momento = (n.º de casos existentes / população total) × *k*; Proporção de Incidência = (n.º de novos casos /população total em risco) × *k* (Mausner e Bahn, 1990; Moon & Gould *et al.*, 2000).

uma delas acumula maior percentagem de população idosa em relação à outra. Situação semelhante pode ocorrer para a taxa de mortalidade por cancro da mama, mais comum em mulheres do que em homens, caso as populações envolvidas manifestem sensíveis desequilíbrios na razão dos sexos.

Tal como reconhece Stone *et al.* (1999), embora as taxas específicas por idade ou por sexo possam ajudar a controlar alguns dos efeitos descritos, a necessidade de ter uma taxa única para determinada área geográfica, válida em termos comparativos, aconselha a "remoção" ou "controlo" destas desigualdades pela metodologia de *padronização de taxas*.

Habitualmente e com recurso a uma população padrão, o processo de padronização pode ser alcançado através da aplicação de dois métodos distintos: método directo e método indirecto de padronização.

O *método directo* requer o acesso aos dados de mortalidade ou morbilidade, discriminados por classe etária, de modo a viabilizar o cálculo de uma taxa específica para cada intervalo etário, cujo resultado ao ser projectado sobre a população padrão permite o cálculo de valores esperados de mortalidade ou morbilidade (Tabela 5.3.1).

Tabela 5.3.1 – Aplicação do método de *padronização directo*

	Concelho			*População Padrão*	
Classe etária (anos) x	População (concelho) P_x	Óbitos por Sida (concelho) O_{ox}	Taxa de mortalidade (concelho) Tm_x	População Padrão P_{px}	Óbitos esperados O_{epx}
15-19	10367	0	0	747346	0
20-24	11633	1	0,0000859	802791	68,96
25-29	11374	2	0,0001758	720773	126,71
30-34	10647	4	0,0003756	691345	259,67
35-39	10642	1	0,0000939	670589	62,97
40-44	10288	5	0,000486	635719	308,96
45-49	10132	3	0,000296	610509	180,71
50-54	8529	3	0,0003517	543386	191,11
55-59	8067	4	0,0004958	528531	262,05
60-64	7913	2	0,0002527	521610	131,81
65+	20964	1	0,0000477	1439942	68,69
Total	*120556*	*26*	-	*7912541*	*1661,64*

Fonte: Adapt. Stone *et al.*, 1999;pp.78

O quociente entre os *óbitos esperados* e a *população padrão* considerada, permite-nos calcular a respectiva taxa de mortalidade, ajustada para a idade, segundo a formulação apresentada por Stone *et al.* (1999; pp. 78):

$$T_{mad} = \frac{\sum O_{epx}}{\sum p_{px}} \times k$$

T_{mad} =Taxa de Mortalidade Ajustada para a Idade (*método directo*),

onde:

O_{ox} = óbitos observados no concelho por faixa etária;

P_x = população do concelho por faixa etária;

$Tm_x = O_{ox} / P_x$; taxa de mortalidade por faixa etária no concelho;

P_{px} = população padrão por faixa etária;

$O_{epx} = Tm_x \times p_{px}$; óbitos esperados na população padrão por faixa etária.

Apesar das vantagens reconhecidas à aplicação do método directo, nomeadamente a comparação das taxas padronizadas entre si, a sua aplicabilidade pode ser inviabilizada pela ausência de informação detalhada sobre o número de óbitos (casos) por classe etária, sendo que, frequentemente, apenas dispomos de informação sobre o total de eventos (casos ou óbitos) observados na população estudada. Outra situação que desaconselha a aplicação deste método prende-se com o baixo número de eventos que podem ser registados em áreas com população reduzida, o que naturalmente retira confiança às taxas específicas calculadas para cada classe. Uma vez que vamos projectar o valor das diferentes taxas de mortalidade (Tm_x) sobre a população padrão (P_{px}), de modo a obter o número de óbitos esperados, o resultado final está, por esta via, sujeito a grande variabilidade, tal como é característico dos pequenos números (Stone *et al.*,1999; Friedman, 1994).

Nestes casos, aplica-se a *padronização pelo método indirecto*, através do recurso a taxas específicas por idades, calculadas para a população padrão, sendo posteriormente projectadas sobre a população em estudo para apuramento dos óbitos ou casos esperados (Tabela 5.3.2). Ou seja, «este processo equivale a perguntar qual seria o número de óbitos [esperados] na população [em estudo] se as pessoas dessa população estivessem a morrer à mesma taxa [taxa específica por idades] a que morrem as pessoas na população padrão.» (Mausner e Bahn, 1990; pp. 489).

Tabela 5.3.2 - Aplicação do método de *padronização indirecto*

	População Padrão			*Concelho*	
Classe etária (anos) *x*	População Padrão P_{px}	Óbitos por Sida O_{px}	Taxa de mortalidade T_{px}	População Concelho P_x	Óbitos esperados O_{ex}
15-19	747346	64	0,000086	10367	0,887792
20-24	802791	648	0,000807	11633	9,389971
25-29	720773	1501	0,002082	11374	23,6862
30-34	691345	1663	0,002405	10647	25,61089
35-39	670589	1230	0,001834	10642	19,51965
40-44	635719	739	0,001162	10288	11,95942
45-49	610509	464	0,000760	10132	7,700538
50-54	543386	338	0,000622	8529	5,305256
55-59	528531	261	0,000494	8067	3,983658
60-64	521610	196	0,000376	7913	2,973386
65+	1439942	244	0,000169	20964	3,552376
Total	*7912541*	*7348*	-	*120556*	*114,5691*

Fonte: Adapt. Stone *et al.*, 1999; pp.79

Também neste caso e para aplicação da *metodologia indirecta*, os procedimentos seguiram a fórmula apresentada:

$$T_{mai} = \left(\frac{O_o}{O_e} \times T_p \right) \times k$$

T_{mai} = Taxa de Mortalidade Ajustada para a Idade (*método indirecto*),
onde:

O_o = total de óbitos observados no concelho;

$O_e = \sum o_{ex}$; total de óbitos esperados no concelho;

$T_p = \frac{\sum o_{px}}{\sum p_{px}}$; taxa bruta de mortalidade da população padrão;

$O_{ex} = T_{px} \times p_x$; óbitos esperados no concelho por faixa etária;

P_x = população do concelho por faixa; T_{px} = taxa de mortalidade por faixa etária da população padrão;

O_{px} = óbitos observados na população padrão por faixa etária;

P_{px} = população padrão por faixa etária.

(adapt. Stone *et al.*, 1999; pp.79)

O autor traz ainda à nossa atenção aquilo que designa por «variante da taxa padronizada por idade pelo método indirecto» - a *Razão Padronizada de Mortalidade* ou *Morbilidade* (RPM[4]), que Mausner e Bahn (1990) também denominam por *Índice Comparativo de Mortalidade*, verificando-se uma frequente utilização ao nível do mapeamento de

[4] RPM – deriva da designação inglesa de *Standardized Mortality Ratio* (SMR).

mortalidade, graças a uma interpretação quase intuitiva que esta razão proporciona (Goldman e Brender, 2000):

$$Razão\ Padronizada\ de\ Mortalidade = \frac{O_o}{O_e} \times 100$$

onde:

O_o = óbitos observados;

O_e = óbitos esperados.

Esta proporção, expressa habitualmente os seus valores em percentagem, sendo que 100 indica o valor esperado em relação à população padrão a partir da qual os restantes valores são referenciados[5].

Embora o método indirecto seja mais facilmente exequível, nomeadamente quando existe pouco detalhe de informação sobre a população estudada, sendo apenas fornecidos os valores totais de um determinado evento, para além de suportar melhor as distorções que podem ser geradas pelo uso de "pequenos números", são diversos os autores que sublinham a incapacidade do método indirecto em corrigir diferenças significativamente marcadas na composição das populações, pelo que Stone *et al.* (1999; pp. 80) adverte:

> «Com o método indirecto não se podem comparar múltiplas taxas. Só a taxa ajustada [para a idade] pode ser comparada com a taxa padrão para uma razão observada ou esperada.»

Na opinião de Friedman (1994; pp. 201), este constitui o principal *handicap* da metodologia indirecta, o que levou muitos epidemiologistas a olharem este procedimento com desfavor, particularmente quando existem *diferenças profundas* entre composições populacionais.

Uma revisão mais detalhada da literatura internacional permite-nos perceber que são muitos e diversos os argumentos esgrimidos sobre este tema, estando ainda longe de constituir matéria consensual. Os autores L. Pickle e A. White (1995), num artigo publicado na conceituada revista *Statistics of Medicine*, tecem veementes considerações sobre o que consideram «a inadequação do metodologia indirecta», especialmente quando aplicada na construção de mapas de mortalidade. Embora reconheçam que proeminentes estatístas aplicam dados calculados por este método na produção de atlas nacionais e internacionais de mortalidade e morbilidade - pressupondo que a

[5] Ex.: uma RPM de 95% deve ser interpretada como sendo uma taxa 5% inferior à verificada na população padrão (Stone *et al.*, 1999;pp. 80).

comparação do valor de cada área deve apenas ser feita com o valor padrão, argumentam que o processo de comparação visual é intrínseco à observação de um mapa, pelo que esta regra tende a ser facilmente violada.

Posteriormente, e de uma forma não menos categórica, A. Goldman e J. Brender (2000) publicam na mesma revista um artigo onde apresentam e discutem conclusões muito diversas das referenciadas no artigo de L. Pickle e A. White (1995).

Após sumariarem as divergências verificadas no seio da epidemiologia sobre a validade de um e de outro método, os autores sublinham que o ajustamento indirecto é preferível quando o número de óbitos na população estudada é muito baixo, capaz de originar taxas específicas instáveis para a idade, sublinhando, uma vez mais, a forte divergência que existe entre os autores neste ponto:

> «Uma revisão bibliográfica sobre textos clássicos em epidemiologia demonstrou a dúvida e o desacordo que existe no seio dos epidemiologistas acerca da validade do método indirecto. Das sete obras consultadas, apenas uma não expressava desacordo sobre a validade dos ajustamentos indirectos [3] [6], três continham advertências directas contra o uso de ajustes indirectos na comparação de diferentes populações [1,5,7]. Duas referências demonstravam desacordo mas expunham a validade do ajustamento indirecto, não citando, na realidade, quaisquer referências ou dados que suportassem as suas pretensões [2,6].» (Goldman e Brender, 2000; pp. 1082).

Por este motivo os autores desenharam um estudo comparativo prosseguindo um e outro método, utilizando inclusivamente um modelo proposto por Rothman (1986), a fim de executarem um vasto conjunto de simulações. Posteriormente, quando da discussão dos resultados alcançados permitiram demonstrar que o ajustamento indirecto [SMRs][7] e o

---

[6][3] Cates w, Williamson GD (1994) Descriptive epidemiology: analysing and interpreting surveillance data. In *Principles and Pratice of Public Health Surveillance*, Teutsch SM, Churchill RE (eds). Oxford University press: Oxford, 96 – 135.

[1] Rothman, Kj.(1986) *Modern Epidemiology*. Little Brown, and Company: Boston, 45-49.

[5] Hennekens CH, Buring JE (1987) *Epidemiology in Medicine*. Little Brown, and Company: Boston, 82-85.

[7] Rothman, KJ, Greenland S (1998) *Modern Epidemiology*. Lippincott-raven: Philadelphia, 234-235, 262.

[2] Selvin S.(1996) *Monographs in epidemiology and Biostatistics. Vol. 25. Statistical Analysis of Epidemiological Data*. Oxford University press: Oxford, 36-40.

[6]Kahn, HA, Sempos CT (1989) *Monographs in epidemiology and Biostatistics. Vol. 12. Statistical Methods in Epidemiology.* Oxford University press: Oxford, 95-98.

[7] Os autores advertem que tomam *SMRs* como *ajustamentos indirectos* e *SRRs* (*Standardized Rate Ratios*) como *ajustamentos directos*.

ajustamento directo [SRRs] detêm similar utilidade na análise dos dados de saúde pública. As magnitudes absolutas do ajuste indirecto e ajuste directo encontradas, foram similares para cada um dos cenários testados. O ajuste indirecto sofreu algumas modificações com as alterações na distribuição etária da população em estudo, mas as mudanças verificadas não foram suficientemente fortes para alterar a interpretação da comparação, mesmo perante cenários extremos aplicados (Goldman e Brender, 2000).
Pelo exposto, e uma vez que não existe uma validade indiscutível de um método sobre outro, porque razão o ajustamento pelo método directo é relativamente menos frequente? Para Goldman e Brender (2000, pp. 1087), o privilegiar da metodologia indirecta em cenários de saúde pública deriva, fundamentalmente, da interpretação quase intuitiva que os resultados desta metodologia proporcionam, nomeadamente a grande facilidade comparativa e explicativa que o método encerra:

> «Uma vantagem clara do ajuste indirecto em cenários de saúde pública é a sua interpretação mais intuitiva, a grande facilidade com que pode ser explicado e comparado. O conceito do número de óbitos esperados e do número de óbitos observados, e a razão entre os dois (SMR) é facilmente assimilado pelo público e pela *media*.
> (…) Além do mais [no método directo], o estudo com pequenas populações pode não alcançar estabilidade nas taxas específicas para a idade. Nestas situações a fiabilidade do ajuste directo é difícil de conseguir. (…)».

Pelo exposto, os autores concluem assertivamente sobre esta polémica questão, especificando que o ajustamento indirecto detém utilidade similar ao ajustamento directo na análise de dados de saúde pública e pode ser inclusivamente utilizado para comparar diferentes áreas geográficas (Goldman e Brender, 2000; pp. 1087).

5.4 - O problema das pequenas áreas e dos pequenos números: aplicação do estimador Bayesiano empírico

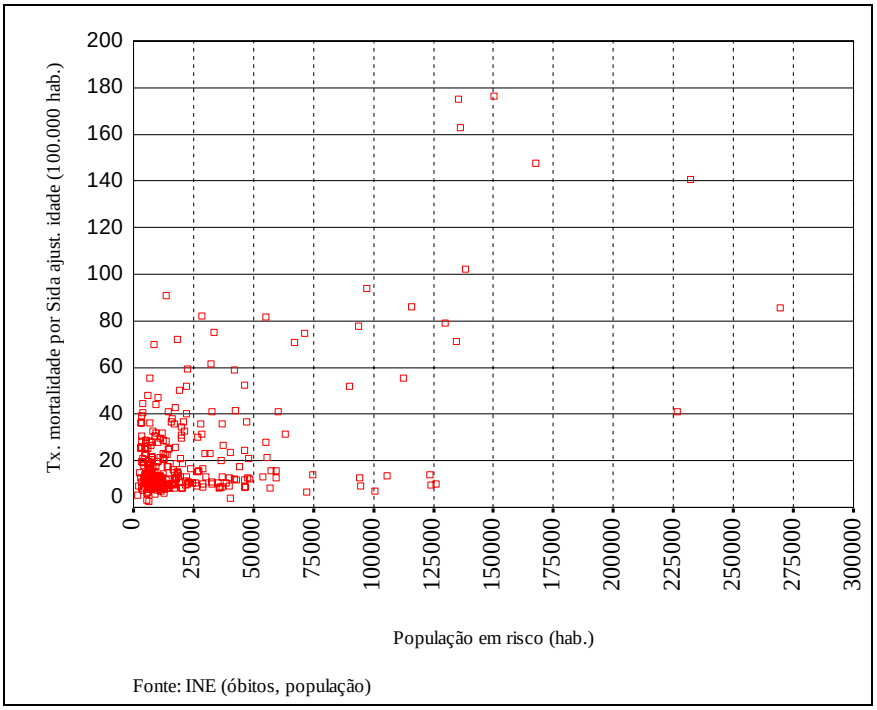
Como verificámos anteriormente, perante a frequente necessidade de comparar indicadores de eventos que atingem diferentes populações no espaço e reconhecendo-se a interferência de factores demográficos no processo, tais como a idade e o sexo, os diversos autores consultados foram unânimes na recomendação da aplicação de um

processo de padronização, com o objectivo de “remover” as diferenças na estrutura populacional, divergindo apenas sobre a validade comparativa final, alcançada pela aplicação de diferentes métodos.

De acordo com Assunção (2001) e Gatrell (2002), quando o risco varia em função das variáveis demográficas apontadas, as diferentes técnicas de padronização revelar-se-ão úteis no “controlo” das diferenças de composição populacional mas insuficientes para obviar à grande variabilidade que pode afectar “eventos raros” numa dada área, num determinado intervalo de tempo. No caso de investigarmos áreas com um universo populacional reduzido, o risco de encontrarmos valores de taxas extremos associados a estas áreas é elevado, originando uma «flutuação aleatória», sobretudo se se considerar a taxa estimada como o resultado da divisão dos eventos observados sobre a população exposta ao risco num determinado intervalo de tempo, onde pequenas diferenças nos valores observados produzirão grandes variações nos valores finais das taxas calculadas para as diferentes áreas (Bailey e Gatrell, 1995; Assunção, 2001).

Um exemplo característico da situação atrás descrita, ocorreu no desenvolvimento do estudo de caso, onde se averiguou a distribuição espacial de óbitos por causa de SIDA para Portugal Continental. Apesar dos eventos se apresentarem agrupados por quinquénio (1991-1995; 1996-2000), de modo a contrariar a flutuação aleatória já mencionada, verificou-se o «efeito de funil» mencionado por Câmara *et al.* [*prelo*], quando comparamos as taxas medidas por cada concelho com a respectiva população em risco (Fig. 5.4.1). Nos concelhos com menores efectivos populacionais (< 25.000 hab.), a variação alcançada oscila entre 0 e 90 óbitos por 100.000 habitantes, transmitindo-nos uma noção de risco pouco consistente, mercê da elevada flutuação dos valores alcançados.

Figura 5.4.1 – Variação da Taxa de Mortalidade por Sida Ajustada para a Idade (1996 – 2000) em função da População em Risco (excepto concelho de Lisboa)



Neste contexto, Assunção (2001;pp. 37) trás à nossa atenção um artigo clássico de Choynowsky (1959) onde o autor trata o problema da oscilação dos pequenos números em pequenas áreas de um modo particularmente pedagógico, a propósito da ocorrência de tumores cerebrais verificados em diversos condados poloneses.

Após o mapeamento das taxas para os sessenta condados do sul da Polónia, Choynowsky constatou que algumas taxas eram muito altas ou muito baixas em comparação com a média da área total, igual a 5,17 por 100 mil habitantes. Procurando compreender essas surpreendentes irregularidades geográficas, o autor observa que os condados que se desviavam muito da média, sem qualquer explicação aparente (qualidade de cuidados médicos, diferenças na composição etária, entre outras), tinham populações pequenas. Por este motivo, uma pequena diferença nas frequências absolutas criava uma diferença substancial nas taxas, evidenciando que a causa dos desvios poderia muito bem ser atribuída às variações amostrais.

Seguindo a descrição de Assunção (2001), verificamos que à menor área de análise, o condado de Lesko, correspondia a menor população e, simultaneamente, a taxa mais elevada:

> «A ocorrência de 2 casos nesse condado é responsável pela taxa de 11,77 por 100 mil. Caso tivesse ocorrido apenas 1 caso, essa taxa baixaria para 5,9 por 100 mil, um valor consistente com as taxas dos outros condados. Por outro lado, a ocorrência de 3 casos faria a taxa pular para 17,7 por 100 mil.».

O efeito drástico que a adição e subtracção de poucos casos acarretava nas taxas calculadas não se verificava nos condados com grandes populações.

Prosseguindo este tipo de reflexão em torno dos pequenos números, Rodrigues (1993), a propósito do estudo geográfico de localizações tumorais para Portugal, descreve o mesmo fenómeno alcançando conclusões semelhantes:

> «(…) o indicador usualmente utilizado para a comparação das taxas de incidência e mortalidade é a RPI (razão padronizada de incidência) ou RPM; no entanto, mesmo este indicador, apesar de possuir um menor erro padrão do que a taxa padronizada pelo método directo, acarreta desvantagens ao ser utilizado no estudo de pequenas áreas, de patologias raras cujo número de casos (óbitos) observados é escasso e de unidades de análise com grandes diferenças de efectivos populacionais (…) Deste modo, a representação geográfica destas RPI e/ou RPM "brutas" poderá ser dominada por valores extremos, provavelmente, na sua maioria, com origem nestas unidades de análise com poucos eventos observados; por outro lado, ela poderá não reflectir uma eventual autocorrelação espacial dos factores de risco igualmente prevalentes em áreas vizinhas.» (Rodrigues, 1993;pp. 85).

O autor adverte ainda para o facto de que a apreensão cartográfica do fenómeno através do cálculo e representação de significância de testes será igualmente inadequada, já que esta metodologia evidenciará apenas pequenas diferenças baseadas em efectivos consideráveis.

Para além do estudo de Rodrigues (1993), de um modo sistemático Bailey e Gatrell (1995), Assunção (2001) e Gatrell (2002), sugerem um conjunto de possíveis soluções, aplicadas a investigações sociais e epidemiológicas, capazes de minimizarem ou mesmo corrigirem os efeitos anteriormente descritos, quase sempre presentes quando se investigam "eventos raros" ocorridos em populações de pequena dimensão. A totalidade dos autores citados, suportados em estudos de localizações tumorais, ou em estudos de mortalidade infantil (Nova Zelândia e Brasil), conclui pela validade da aplicação de um

modelo que produz estimativas ajustadas espacialmente da taxa de incidência ou da taxa de mortalidade, através da aplicação de um *estimador Bayesiano empírico local*.

Gatrell (2002; pp. 53-60) antecipa à aplicação do estimador Bayesiano empírico local, um outro conjunto de possíveis soluções, mais fáceis de simular, mas que se podem revelar menos eficientes na resolução do problema apresentado – variabilidade de pequenos números em pequenas áreas.

Como solução inicial, de aplicação simples e potencialmente correctora da emergência de valores com elevada flutuação observados em áreas com menor população, o autor sugere que a metodologia de investigação contemple o processo de agregação de um determinado número de anos de casos ou óbitos observados – que podem ser agrupados por triénio, quinquénio ou decénio, entre outras, de modo a que se possam alcançar valores mais estáveis para o evento em causa.

Outra hipótese a considerar, pode passar pela modificação da escala de análise, considerando áreas geográficas de maior dimensão e menor escala o que inviabiliza, parcial ou totalmente, a investigação da distribuição geográfica da morbilidade e mortalidade ao nível local.

Por último, sugere o cálculo de intervalos de confiança para os indicadores, ou a construção de mapas de probabilidades, através de uma distribuição de Poisson, o que comporta a dupla desvantagem de assumir o valor de cada área de forma independente e de sustentar a tendência de destacar áreas com maiores efectivos populacionais (Gatrell, 2002;pp. 54-55).

Como alternativa e respeitando os princípios já mencionados, implícitos à autocorrelação espacial, Bailey e Gatrell (1995), Assunção (2001) e Gatrell (2002), sugerem a aplicação do *Estimador Bayesiano Empírico,* como modelo de estimação e ajustamento de taxas, capaz de contornar a flutuação aleatória de valores associados a pequenas populações. A aplicação do estimador Bayesiano empírico considera o princípio básico da análise espacial admitindo, a *priori*, que áreas próximas tendem a ser mais parecidas do que áreas mais distantes. Por este motivo, utiliza o registo dos valores obtidos para uma dada variável nas áreas de vizinhança, como um indicador suficientemente realista para (re)calcular ou "corrigir" os valores verificados em áreas com menores efectivos populacionais e, consequentemente, com maior propensão para a aleatoridade (Bailey e Gatrell, 1995).

De acordo com Bailey e Gatrell (1995), por se tratar de um estimador *empírico* admite-se, a *priori*, igual distribuição da variável aleatória θ_i para todas as áreas. Então a taxa

estimada $\hat{\theta}_i$ terá uma probabilidade de distribuição para cada área de média *m* (o que corresponde à taxa média global se calcularmos $\hat{\theta}_i$ a *nível global*; ou corresponderá à taxa média dos vizinhos se calcularmos $\hat{\theta}_i$ a *nível local*, m_i), e variância ν.

Para estimar θ_i localmente, combina-se a distribuição apriorística (taxa média local dos vizinhos m_i e variância ν) com as taxas observadas t_i (taxa bruta observada, onde o numerador é composto pelo número de eventos observados em cada área *i* e o denominador compreende a população existente e exposta ao risco na mesma área, n_i):

$$\hat{\theta}_i = w_i t_i + (1 - w_i)\, m_i$$

onde:

$$w_i = \frac{\nu}{\nu + \frac{m_i}{n_i}}$$

$$\nu = \frac{\sum_{i=1}^{k} n_i (t_i - m_i)^2}{n} - \frac{m_i}{\bar{n}}$$

$\hat{\theta} i =$ taxa estimada para cada área *i*;

w_i = peso ou factor de ajustamento;

n_i = população da área *i*;

k = n.º de vizinhos;

$n = \sum_{1}^{k} n_i$; total da população dos vizinhos da área *i* (excluindo a área *i*);

$\bar{n} =$ média da população dos vizinhos da área *i*;

O_i = n.º de óbitos da área *i*;

$t_i = \frac{O_i}{n_i}$ taxa bruta na área *i*;

$m_i = \frac{\sum_{1}^{k} O_i}{n}$, taxa média observada nos vizinhos da área *i*.

Por convenção:

Se $\nu_i < 0 \Rightarrow \nu_i = 0$, logo $w_i = 0$ e $\hat{\theta}_i = m_i$

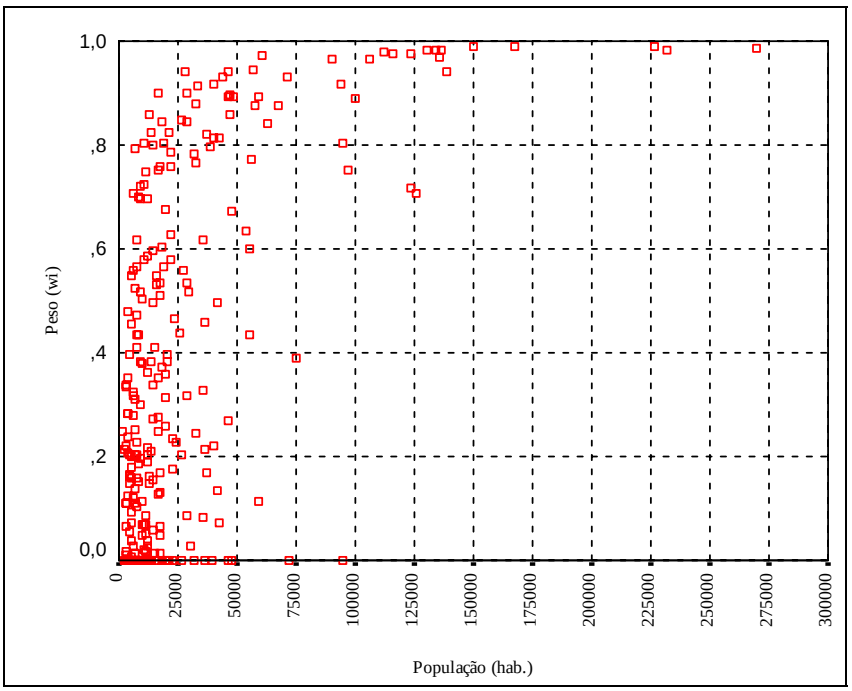
(Bailey e Gatrell; 1995; pp. 304-306; Souza *et al.*, 2001; pp. 476)

O peso w_i varia entre 0 e 1, em função do tamanho da população e da variância ν. Assim w_i aproxima- se de 1 em áreas onde a população contém maiores efectivos e a

"confiança" atribuída à taxa observada é maior, pelo que o valor estimado de $\hat{\theta}_i$ adquire valores iguais ou semelhantes aos observados. Inversamente, em áreas de menor base populacional e sujeitas a flutuações aleatórias de eventos, w_i aproxima-se de 0 e o valor da taxa observada na área *i* é ajustado ou "contraído" em função do valor alcançado pela taxa média dos vizinhos (m_i), (Bailey e Gatrell, 1995).

Estamos pois perante a aplicação de um algoritmo que permite (re)estimar taxas onde o peso das taxas médias dos vizinhos de *i* tende a ser inversamente proporcional ao tamanho da população da área *i*. Tal como se demonstra na figura 5.4.2, no caso da aplicação do estimador Bayesiano empírico local à Taxa de Mortalidade por Sida Ajustada para a Idade em Portugal (continental), os menores valores de w_i são alcançados, maioritariamente, em áreas onde a população < 25.000 habitantes, aproximando-se de 1 à medida que se atingem efectivos populacionais de maior dimensão, neste caso >50.000 habitantes.

Figura 5.4.2 - Variação Peso (w_i) *versus* População (hab.), segundo a aplicação do estimador Bayesiano empírico local aplicado à Taxa de Mortalidade por Sida Ajustada para a Idade (1996 – 2000), (Portugal continental, excepto concelho de Lisboa).



Podemos pois observar que a aplicação do estimador Bayesiano empírico local permite estimar valores mais estáveis para os indicadores de risco calculados – neste caso Taxa de Mortalidade por SIDA Ajustada para a Idade, tal como se demonstra na figura 5.4.3,

onde os valores menos confiáveis das taxas observadas, gerados sob a influência de "pequenos números", em áreas de menor dimensão populacional, são objecto de "suavização", alcançando-se valores mais confiáveis, cuja validade é alcançada em função dos valores médios da variável registados na vizinhança (m_i) e em função do peso (w_i) atribuído à área.

Figura 5.4.3 - Comparação da Taxa de Mortalidade por Sida Ajustada para a Idade *versus* Taxa de Mortalidade por Sida ajustada para a idade estimada pelo *método bayesiano empírico local* (1996 – 2000), Portugal continental.

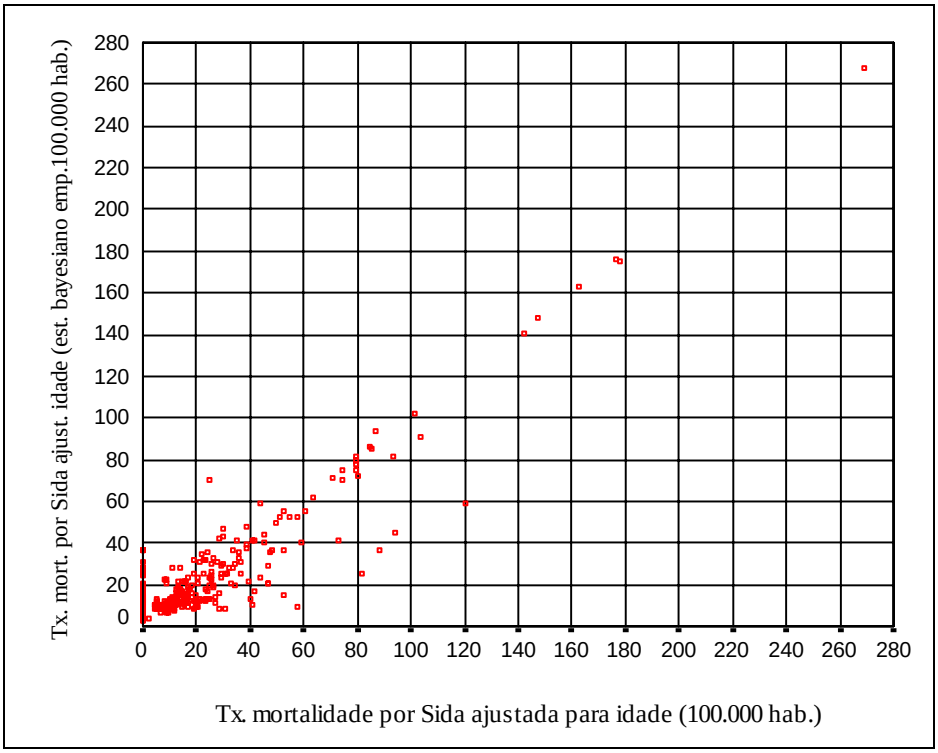


Fig 3 - Comparação da Taxa de Mortalidade por Sida Ajustada para a Idade *versus* Taxa de Mortalidade por Sida ajustada para a idade estimada pelo *método bayesiano empírico local* (1996 – 2000), Portugal continental.

5.5 - Utilização de estimadores de intensidade

Frequentes vezes, os processos de investigação sócio-demográficos e epidemiológicos beneficiam da utilização de ferramentas que lhes permitem estimar a intensidade de um fenómeno, medido no espaço de forma discreta, interpolando valores para todos os pontos da área em estudo. Uma vez que, maioritariamente, a investigação desenvolvida sobre este tipo de dados recai sobre atributos associados a áreas, confinados por linhas

de fronteira e vizinhança, gerados quase sempre por conveniência administrativa, a utilização de *estimadores de intensidade* permite contornar a situação descrita, como que "dissolvendo os limites das áreas", tomando o espaço em contínuo.

Para a geração das superfícies contínuas, é necessário utilizar métodos de interpolação espacial, que partem de um conjunto de pontos amostrais (verdadeiros ou conceptuais). Assim, a interpolação constitui-se num método adequado para determinar valores em pontos desconhecidos, partindo de um conjunto de pontos amostrais situados em localizações conhecidas, baseando-se na premissa de que pontos mais próximos no espaço têm uma maior probabilidade de ter valores similares do que pontos mais distantes.

O processo de predição de valores efectua-se através do ajuste de uma função matemática que melhor descreve o comportamento do fenómeno investigado. A este nível, são diversos os modelos matemáticos usados na metodologia de interpolação pelo que, em Geografia Humana, os mais referenciados baseiam-se na predição de valores a partir do inverso da distância dos pontos amostrais ao ponto a estimar.

Deve também notar-se que os métodos de interpolação podem ser *globais* ou l*ocais*, sendo que os primeiros consideram para a estimação de um valor a totalidade dos pontos amostrais existentes na região. Contrariamente, os estimadores locais consideram apenas os pontos amostrais presentes numa área de influência, cuja definição constitui um dos parâmetros da função matemática, relevando o critério do investigador, baseado no conhecimento que o mesmo detém sobre a distribuição do fenómeno estudado.

De um modo geral, os métodos de interpolação de pontos seguem uma sequência de procedimentos que se origina com a sobreposição de uma grade regular, que divide a área de estudo em células (*pixels*), cujo tamanho deve ser determinado quando da planificação dos procedimentos de interpolação. Quanto maiores forem as células menor é a resolução das superfícies pelo que, pequenas variações do fenómeno investigado podem passar despercebidas. Neste sentido, o tamanho das células deve ser ajustado em função dos dados que estão a ser analisados (regra geral este processo envolve diversas tentativas).

O segundo passo da interpolação envolve a estimativa de valores para cada célula partindo dos pontos amostrais. Se o método for global, todos os pontos amostrais existentes na região de estudo contribuem para a estimativa. No caso de se utilizarem métodos locais, os pontos amostrais considerados em cada estimativa (z_i), são

determinados em função de um raio de influência ( ), centrado no ponto a estimar ($\hat{z}_s$), i.e. só contribuem para a estimativa os pontos amostrais que estiverem numa posição (s_i), em relação à posição (s) do ponto a estimar, de tal modo que a distância entre os dois seja inferior ao raio de influência $(s - s_i \leq \tau)$.

O terceiro passo do processo envolve a aplicação do método de interpolação escolhido, onde se procede à estimativa de valores para cada célula da grade, respeitando um princípio comum: pontos amostrais mais distantes do ponto a estimar contribuem menos para o valor a estimar do que pontos amostrais mais próximos.

Neste processo, é relativamente frequente a citação do estimador *Kernel*, (Bailey e Gatrell, 1995; Robinson, 1998; Gatrell, 2002), mesmo quando é aplicado na estimação de modelos contínuos e que se traduz pela seguinte fórmula:

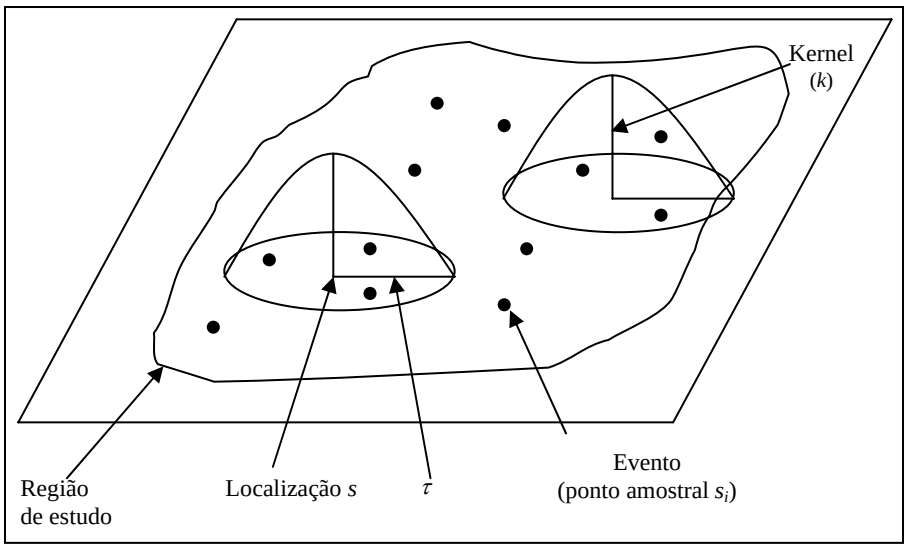
$$\hat{z}_s = \frac{\sum_{i=1}^{n} k\left(\frac{s - s_i}{\tau}\right) \times z_i}{\sum_{i=1}^{n} k\left(\frac{s - s_i}{\tau}\right)}; s - s_i \leq \tau$$

onde:

$\hat{z}_s$ = valor a estimar ou predizer;
τ = raio de influência;
z_i = valor da variável considerada na posição *i*;
k = função de kernel (interpolador não-paramétrico);
s = localização do valor a estimar;
s_i = localização da variável.

(Bailey e Gatrell, 1995)

Figura 5.5.1 - Esquema de funcionamento de um estimador de intensidade [Kernel], aplicado a uma distribuição pontual



Fonte: Adapt. Bailey e Gatrell, 1995; pp. 86

Existem, todavia, outros estimadores que utilizam no processo de interpolação espacial algoritmos diversos, onde a função de estimação não é tão complexa como a de Kernel, considerando igualmente os valores estimados para cada ponto em função da distribuição de um peso, inversamente proporcional à distância entre os pontos amostrais e os pontos a estimar. Por este motivo são vulgarmente referenciados como *interpoladores de distância inversa* (IDI) que, tal como o nome indica, pressupõem que os valores mais próximos dos pontos a interpolar têm maior influência do que os valores mais afastados. Assim, a maioria das ferramentas informáticas básicas que utilizam interpoladores de distância inversa seguem a seguinte fórmula:

$$\hat{z}(s_o) = \sum_{i=1}^{n} w_i z(s_i)$$

onde:

$\hat{z}(s_o)$ = valor a predizer na posição S_0;

$z(s_i)$= valor observado na localização S_i;

w_i = peso associado a cada par de coordenadas a utilizar (varia inversamente com a distância);

n = número de pontos amostrais

Neste contexto, deve ainda considerar-se a existência de dois tipos distintos de interpoladores: aqueles cuja técnica de interpolação procede à distribuição dos valores estimados pelas células da matriz, mantendo inalterado o valor adstrito ao ponto amostral, pelo que se denominam por *interpoladores exactos;* ou aqueles que predizem os valores para as células da matriz, modificando o valor da variável inicialmente associada ao ponto amostral e que se designam por *interpoladores inexactos* (Fig.5.5.2).

Figura 5.5.2 – Geração de uma superfície de densidade a partir de um *Interpolador de Distância Inversa* para a *Estimativa Bayesina Empírica de mortalidade por causa VIH/SIDA* (1996 – 2000)