MINERAÇÃO DE DADOS ESPACIAIS: A BUSCA DE PEPITAS DE OURO

O levantamento de dados em campo representa a tarefa de maior custo em aplicações de Geoprocessamento que usam dados sócio-econômicos, como Geonegócios, Políticas Públicas, Saúde, Educação e Planejamento Urbano. Por este motivo, é preciso extrair o máximo possível dos dados que dispomos, atividade comparável ao trabalho do minerador, que através de uma "bateia" (o GIS) busca encontrar "pepitas de ouro" (padrões e relacionamentos existentes, mas que não são evidentes).

Tomemos um exemplo concreto: os dados do "Mapa de Exclusão/Inclusão Social na Cidade de São Paulo", resultado de uma pesquisa multi-institucional liderada pela profa. Aldaíza Sposati da PUC/SP (que também é vereadora pelo PT na capital). A cidade de São Paulo foi dividida em 96 distritos, cujos dados incluem: população, número de idosos, renda por faixa salarial, total dos domicílios, casas sem esgoto, número de analfabetos, número de crianças na escola.

Em seu excelente artigo "Síndrome de Óculos para Perto", publicado na InfoGeo nº. 7, nosso colega Francisco Aranha analisou estes dados, utilizando o percentual dos moradores com mais de 70 anos como indicativo da qualidade de vida. Constatou uma grande variação entre o centro (onde a proporção de idosos é de 6%) com a periferia (com apenas 1%). Neste artigo, utilizaremos este mesmo indicador, buscando obter informações adicionais.

Assim, queremos saber: *Existem regiões com variações extremas de longevidade, que indicam casos críticos de exclusão ou inclusão social ?* Inicialmente, calculamos a média e o desvio padrão do percentual de idosos e plotamos o mapa de distribuição espacial (figura 1), que mostra duas regiões com valores muito acima da média (mais de 2 desvios padrões): Jardim Paulista, Consolação e Lapa, todos com mais de 7,5% de idosos.

Este primeiro resultado mostra os valores extremos, mas não fornece indicações sobre a dinâmica espacial do processo. Para ir além, precisamos usar o princípio básico do Geoprocessamento: "o lugar faz a diferença". Para

tanto, calculamos a *média local* do número de idosos para cada distrito (obtida a partir da vizinhança de cada região) e comparamos o resultado com o valor do atributo para o distrito. Se a diferença for muito pequena, podemos estar diante de regiões com dinâmica própria, fortemente relacionadas, denotadas pelo jargão "aglomerados espaciais", e que indicam "bolsões" de exclusão ou inclusão social.

O resultado do indicador baseado na média local está mostrado na figura 2, onde estão destacadas as regiões que formam aglomerados espaciais, com base em critério de significância estatística (probabilidade de ocorrência aleatória menor que 5%). Percebemos que na zona Leste e na zona Sul de São Paulo há regiões críticas, onde o agravamento das condições sociais resulta numa degradação do nível de vida ainda mais séria que poderiamos esperar. Por exemplo, no caso da zona Sul, os distritos de Jardim Ângela (0,9% de idosos), Parelheiros (1,1%), Capão Redondo (1,2%), Jardim São Luís (1,2%), Cidade Dutra (1,54%) e Grajaú (0,85%) formam um "bolsão" bem-definido de exclusão social.

Para continuar nossa mineração, vamos tentar estabelecer uma correspondência entre as variáveis. Fazemos a hipótese que a quantidade de idosos é condicionada pelas condições de higiene (domicílios com rede de esgoto) e de renda (chefes de família que ganham mais de 20 salários mínimos). Com uma regressão linear tradicional, não temos muito sucesso: as condições de higiene e renda explicam apenas 30% da variação do numero de idosos.

Estariam higiene e renda tão pouco relacionadas com a qualidade de vida ? Mais uma vez, precisamos levar o espaço em conta, pois São Paulo apresenta grandes diferenças entre o centro e a periferia. Uma conjectura plausível é supor que a relação entre as variáveis funcione de forma diferente no centro e na periferia. Para testar esta idéia, dividimos a cidade em três grupos: o centro, a periferia, e a transição centro-periferia (figura 3) e realizamos uma regressão linear separada para cada um deles, buscando explicar a variação do número de idosos a partir dos dados de higiene e renda.

Como resultado, descobrimos que a população da terceira idade está fortemente relacionada com higiene e renda, mas que a forma de interação é diferente para cada parte da cidade. Ao combinar as três estimativas, verificamos que as condições de higiene e de renda explicam mais de 85% da distribuição de idosos. A figura 4 ilustra os erros encontrados nesta estimativa, mostrando que a relação de higiene e renda com longevidade é ainda mais forte na periferia da cidade (onde os erros foram menores) que no centro.

Surpreendente ? Nem tanto, se consideramos o que diz o Prêmio Nobel Amartya Sen, economista indiano e grande estudioso dos problemas de exclusão social. Ele nos ensina que a definição de pobreza deve incluir não apenas a renda, mas também a capacidade de converter esta renda em qualidade de vida. Na periferia, onde a renda é muito baixa, uma pequena melhora nas condições de higiene tem reflexos imediatos em questões como longevidade.

Em resumo, a combinação de técnicas de estatística convencional e análise espacial fornece ao usuário meios importantes de "garimpar" padrões e relações em seus dados, e calçar suas hipóteses em sólidos fundamentos quantitativos.

O autor agradece ao prof. Francisco Aranha pela cessão dos dados georeferenciados de São Paulo, e à profa. e vereadora Aldaiza Sposati pela permissão de utiliza-los. Os dados apresentados estão descritos no livro "Mapa da Exclusão/Inclusão da Cidade de São Paulo" (editora EDUC). Material complementar sobre o presente artigo pode ser obtido em minha homepage (www.dpi.inpe.br/gilberto/infogeo).

Gilberto Câmara (www.dpi.inpe.br/gilberto) é coordenador de Pesquisa e Desennolvimento em Geoprocessamento do INPE, e é um dos responsáveis pelo desenvolvimento do sistema SPRING.

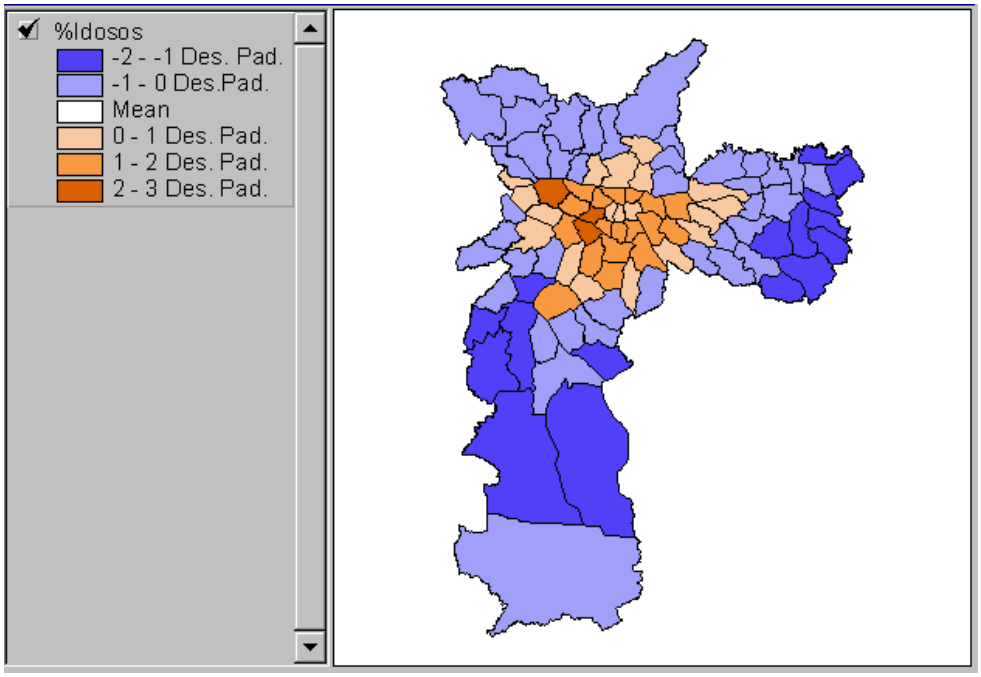

Figura 1 – Distribuição estatística da percentagem de idosos por distritos de São Paulo (laranja: acima da média; azul: abaixo da média). Média dos dados = 3,43% e desvio padrão = 2,04.

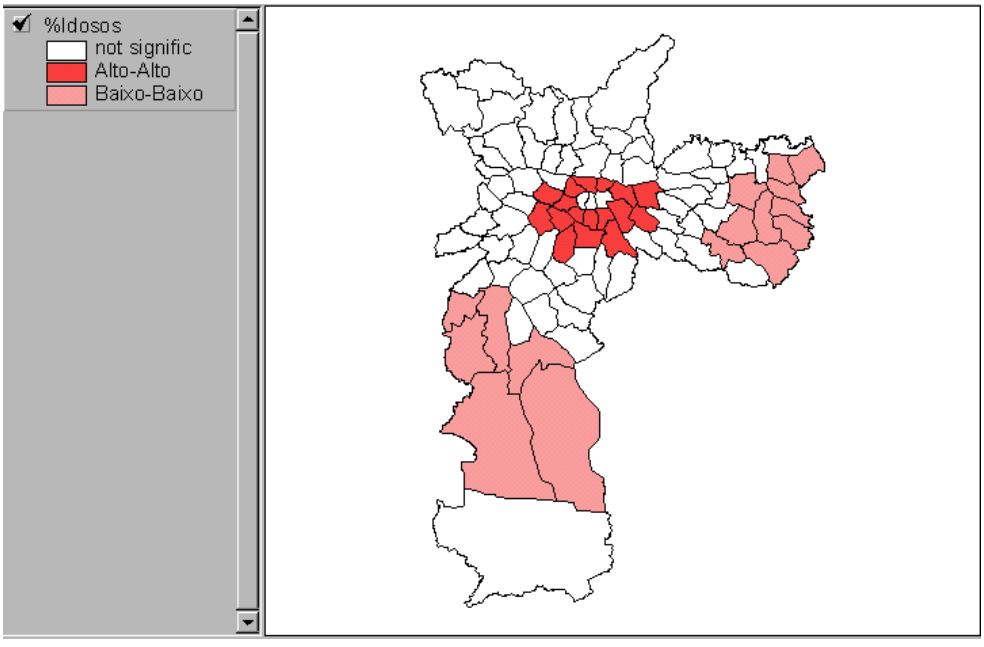

Figura 2 – Aglomerados espaciais na cidade de São Paulo, considerando o percentual de idosos com indicador. Vermelho: "bolsões" de inclusão social; rosa: "bolsões"de exclusão social.

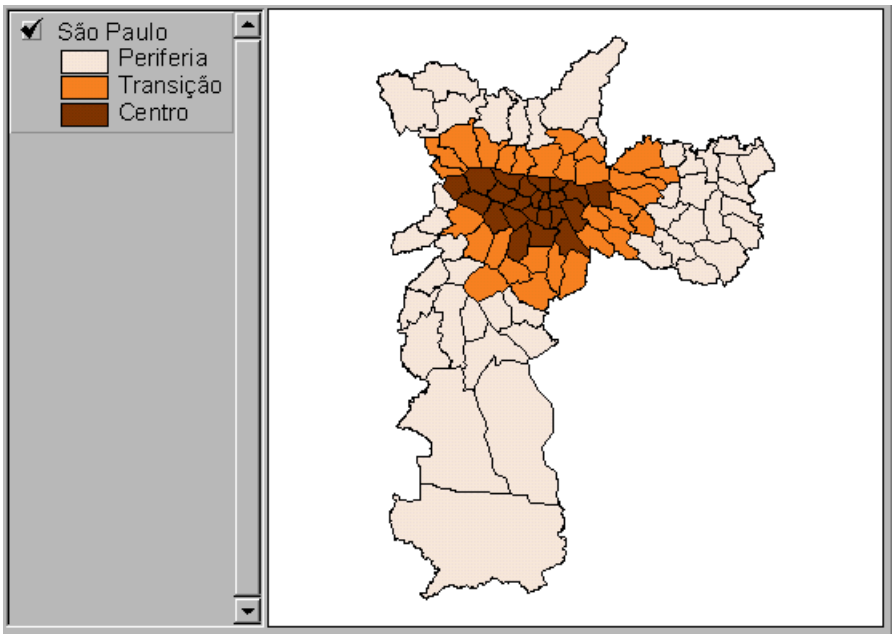

Figura 3 - Divisão da cidade de São Paulo em regiões, para diferenciar o centro da periferia.

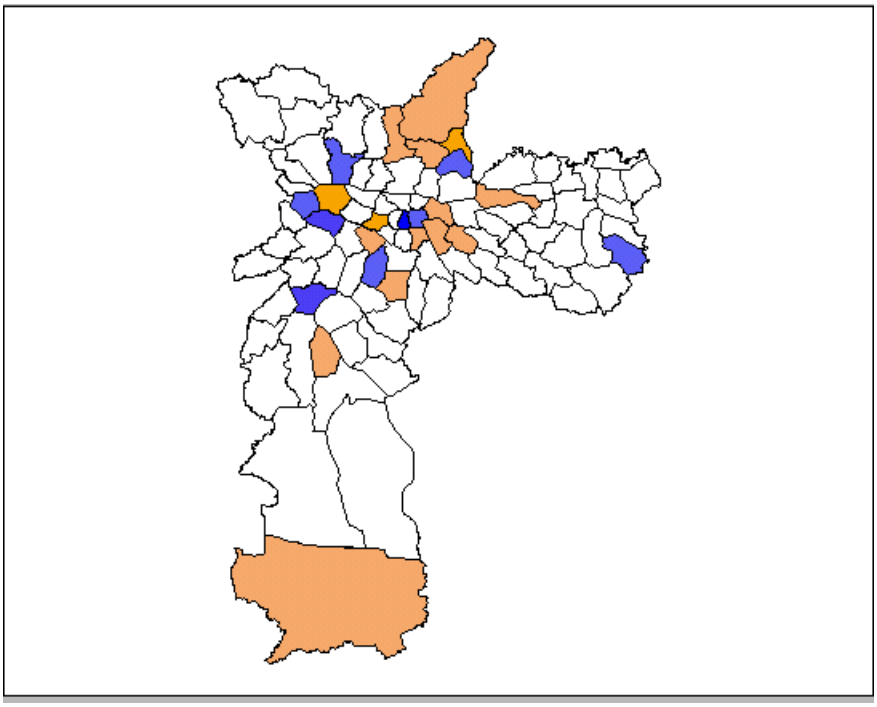

Figura 4 - Resíduos da estimação da proporção de idosos através da rede de esgotos e da renda para a cidade de São Paulo. Em azul, locais onde os valores foram superestimados em mais de 1,5%; em amarelo, distritos onde os valores foram subestimados em mais de 1,5%. Em branco, distritos onde a estimativa teve erro absoluto menor que 1,5%.