平成１２年度厚生科学研究費補助金

統計情報高度利用総合研究事業報告書（課題番号 H11-統計-004）

人口動態統計指標のベイズ推定と地域集積性の評価に関する研究

主任研究者　丹後俊郎（国立公衆衛生院付属図書館長）

分担研究者

山岡和枝　（帝京大学法学部助教授）

今井　淳　（高知県健康福祉部薬務衛生課）

２００１年３月

平成１２年度厚生科学研究費補助金

統計情報高度利用総合研究事業報告書

人口動態統計指標のベイズ推定と地域集積性の評価に関する研究

目次

総括報告

厚生科学研究費補助金（統計情報高度利用総合研究事業）
総括研究報告書

人口動態統計指標のベイズ推定と地域集積性の評価に関する研究

主任研究者　丹後俊郎　国立公衆衛生院付属図書館

研究要旨：本年度は、最終年度として二つの分担研究を行った。まず、人口動態統計指標のベイズ推定と地域集積性の方法論に関する研究においては、疾病地図における地域集積性を検出する二つの方法（Tangoの検定 のKulldorff の検定）について様々な集積モデルによるシミュレーションによる比較を検討するとともにその実用性を検討した。疾病の集積モデルとしては、ある地域を中心とした集積モデル(hot spot cluster) と幾つかの地域の連鎖における集積モデル(chain cluster)、人口の大きさを考慮した地域モデルとして過疎地域、中都市、大都市を設定して比較をおこなった。結果は、Tangoの方法では大都市での集積性、chain cluster に対する検出力が高く、Kulldorffの方法は過疎地域での集積性にたいする検出力が大きかった。また、Kulldorffの方法は広域的な施策体系の企画・立案に有用であること、Tangoの方法は、高死亡率地域のみならず低死亡率地域まで検出しているので、地域の疾病状況の全体像が把握できる点で有用であることが検証できた。市区町村別疾病地図の視覚的表示解析システムに関する研究では、昨年度の基本設計にしたがって地域別指標の視覚表示（疾病地図）を高速に実現でき、かつモデルのよさを相対的に評価できる道具としての疾病地図の視覚的表示解析システムソフトをWindows 95/98上で開発することに成功するとともに、その実用性を目的外使用で申請した人口動態統計死亡データで評価した。

分担研究者

山岡和枝（帝京大学法学部助教授）
今井　淳（高知県健康福祉部薬務衛生課）

A. 研究目的

今日公表され利用されている厚生統計指標は、年齢調整死亡率、標準化死亡比などのように市区町村などの地域の「人口の年齢分布の違い」を調整しているものの、「人口の大きさ」までは調整できていない。そのため、これらの指標を利用して数区分に色分けした疾病地図を作成すると、人口の小さい地域の指標のバラツキが大きく、わずかな死亡数の変化が見かけ上の指標を大きく変化させるという不安定性が指摘されている。つまり、人口の相対的に小さい地域には、死亡率の高い、または低い地域が集まり、人口の大きい地域は真ん中の水準、というように本来の疾病地図の目的とは異なった結果が得られてしまう。本研究の第１の目的は、この問題を解決するために、近年、統計学の世界で計算機の発展によってその重要性が再認識され方法論が大きく進歩したベイズ流アプローチを利用して「人口の大きさ」を調整し、より適切な人口動態統計指標を開発しその利用を推進させることを目的とする。本研究の第２の目的は、疾病の地域集積度指標の開発である。限りある予算、資源を効率的に投入して地域ニーズに対応したきめ細かい対策の立案・実施を行うためには、対策が最も必要とされている最優先地区を選定する必要がある。この目的のためには、地域別に推定された疾病指標（ベイズ推定も含め

て）では、大小に並べれば「必ず」最も高い地域が検出されるという意味で不適切であり、「疾病の集積性」を表現する別の指標を導入しなければならない。統計学的手続きとしては、「対象地域における疾病の地域集積性の有無を検定し、有意な集積性が認められた場合にその地域を推定する」方法論である。第３の目的は、これらの新しい指標の普及に向けた市区町村別疾病地図の視覚的表示解析システムを Windows 上で開発することにある。

B. 研究方法

本年度は、以下の２つの分担研究を行った。

１．人口動態統計指標のベイズ推定と地域集積性の方法論に関する研究（分担者　丹後俊郎、今井 淳，山岡和枝）

２．市区町村別疾病地図の視覚的表示解析システムに関する研究（分担者　今井 淳、丹後俊郎）

分担研究１では，今年度は，疾病地図における地域集積性を検出する二つの方法（Tango の検定とKulldorff の検定）について様々な集積モデル，地域による違いを検討するために，(1) New York, Boston, Washington などの都市を中心としたアメリカ東部地域にいて疾病集積モデルに基づくシミュレーションによる比較をKulldorff と共同研究を行った。(2) 日本の平成５年から９年までの人口動態統計死亡データを用いてその実用性を高知県のデータで検討した。アメリカ東部については，ある地域を中心とした集積モデル(hot spot cluster) と幾つかの地域の連鎖における集積モデル(chain cluster) を仮定し、人口の大きさを考慮して過疎地域、中都市、大都市毎の集積モデルを設定して比較をおこなった。

高知県における検討では、高知県内の５３全市町村における平成５年から平成９年までの５年間の人口動態調査（死亡票）を基に、新たに開発した市区町村別疾病地図の視覚的表示解析システム（Disease Mapping System）を使用し、SMR、SMRの経験的ベイズ推定量（$EBSMR$）、Tangoによる集積性の検定及び、Kulldorffによる集積性の検定を行うことにより検討する。

分担研究２では，プログラム開発言語の一つであるMicrosoft Visual Basic を用いて日本全国の県・二次医療圏・市区町村別の人口動態統計指標（死亡統計）に基づく疾病状況を視覚的に表示する地図（疾病地図）を高速に描写し、かつ、疾病地図における地域集積性を検出するための二つの方法（Tangoの検定とKulldorffの検定）を相対的に評価できる疾病地図の視覚的表示解析システム（ソフト）を開発する。

C. 研究結果

１．分担研究１

まず，アメリカ東部におけるシミュレーションの結果は別添英文報告に詳しく述べたが，Tangoによる方法と、Kulldorffの性質の違いがよく結果に現れた。結果は、Tangoの方法では大都市での集積性、chain cluster に対する検出力が高く、Kulldorffの方法は過疎地域，中都市での検出力が高い結果であった。ただ，Kulldorffの方法は集積していない周辺地域も巻き込んだ形での集積地域の推定を行ってしまう危険性が指摘された。高知県の検討では、Tangoの方法は、地域の中で最も相対危険度が高い地域を検出し、これに対しKulldorffの方法は相対危険度が高い地域の広がりを検出する傾向が観察された。また、Tangoの方法では、低死亡率側に有意な地域も検出してくれる点がメリットであると考えられた。一般的に，SMRを階級別に色分けして疾病地図を作製している例が多いが、この場合、ある特定地域の相対的な位置を主観的・視覚的に見分けはつくが、疾病集積性の客観的な判断材料(EBM: Evidence Based on Medicine)には応用できない。SMRのχ^2検定結果はその性質上、人口規模が大きいところが有意となる傾向が強いため、誤った認識を与える可能性がある。また、人口規模を調整した$EBSMR$の場合は、客観的な判断材料にはなるがそれでも情報量が多すぎる。その点、Tangoによる方法と　Kulldorffによる方法では、表現方法にこそ違いはあるものの、情報をよく集約できており、集積性のある地域を客観的、視覚的に

よく表現できている点が評価できた。

2．分担研究2

昨年度に検討したシステムの基本設計にしたがって開発を進めた。地域別指標の視覚表示（疾病地図）を高速に実現でき、かつモデルのよさを相対的に柔軟に評価できる道具としての疾病地図の視覚的表示解析システムソフトをWindows 95/98上で開発することに成功した。また、その実用性を目的外使用で申請した人口動態統計死亡データ（３大死因である「悪性新生物、脳血管系疾患、循環器系疾患」それぞれのいくつかの死因、希少生起事象である「喘息」、大きな地域差はないと考えられる「先天性異常」）で検討した。地図を表示する場合には、境界データが不必要に細かすぎるため、データを間引いて単純化し、かつ、質の高い表示が可能なように地図情報を最適化するとともにデータの抽出・図の表示速度を可能な限り高めることにもある程度成功した。

D. 考察

今年度は研究の終了年度であるが、当初の研究目的をほぼ達成できたと考えている。つまり、疾病対策の企画立案にあたって地域の疾病状況を把握するための各地域の安定した疾病地図として「人口の大きさ」を調整できる「経験ベイズ推定量」が期待どおり精度のよい疾病状況を表現できることを確認し、それを実用化できたこと。更に、疾病の集積性の高い（または低い）地域を容易に把握できる新しい方法を開発し、Kulldorffの方法とともに視覚的表示解析ソフトとして実用化できたことである。もちろん、Windows95/98 上で開発した視覚的表示解析ソフトは完成品としてはまだまだ不十分であるものの、様々な疾病地図を柔軟に表示でき、疾病の地域集積性が検討できる機能がそろっている点で実用性が充分である。真に対策が必要な地域の同定が容易となるばかりでなく効率的かつ有効な対策の企画立案に有用であることを期待したい。コンピュータソフトは希望者に無料で配布して更なる改善を検討したい。

E. 結論

1）本研究で検討した人口動態統計死亡指標の経験ベイズ推定、集積性の検定はこれまでの指標では検討できない優れた特徴があることを人口動態統計データ，シミュレーションなどにより検証した。

2）集積性の評価手法として用いた Tangoによる方法と、Kulldorffの性質の違いがよく結果に現れた。Tangoの方法は大都市の集積性に，過疎地域での集積性の検出には Kulldorffの方法がより適していることが示された。その利用に当たっては，その目的・性質をよく理解して適用することが重要である。

3）また，Tango の方法は、高危険度の中心点を検定しているため理解が得られやすく、種々の調査や行政施策の費用対効果もあがりやすく、かつ高死亡率地域のみならず低死亡率地域まで検出しているので、地域の疾病状況の全体像が把握できる点で大きなメリットがあると考えられる。

4）Windows 95/98/2000/ME/NT(4.0)上で人口動態統計指標（死亡統計）のSMR, 経験Bayesを利用したEBSMRの視覚表示（疾病地図）を高速に描写し、地域集積性を検出するための二つの方法（Tangoの検定とKulldorffの検定）を相対的に評価できる疾病地図の視覚的表示解析システムを完成した。

F. 研究発表

学会発表

1)Tango, T. Extended score tests for focused clustering. The XXth International Biometric Conference, July, Berkeley, California, U.S.A. 2000.7 ; p93

2)中谷実，丹後俊郎 青森県における疾病の地域集積性．第59回日本公衆衛生学会，群馬，2000.10, p829.

分担研究報告

人口動態統計指標のベイズ推定と地域集積性の方法論に関する研究
(統計情報高度利用総合研究事業) 分担研究報告書

研究者　丹後俊郎　　国立公衆衛生院付属図書館長
研究者　今井　淳　　高知県健康福祉部薬務衛生課
研究者　山岡和枝　　帝京大学法学部助教授

研究要旨：疾病地図における地域集積性を検出する二つの方法（Tango の検定とKulldorff の検定）について様々な集積モデルによるシミュレーションによる比較を検討するとともにその実用性を高知県のデータで検討した。疾病の集積モデルとしては、ある地域を中心とした集積モデル(hot spot cluster) と幾つかの地域の連鎖における集積モデル(chain cluster)、人口の大きさを考慮した地域モデルとして過疎地域、中都市、大都市を設定して比較をおこなった。結果は、Tangoの方法では大都市での集積性、chain cluster に対する検出力が高く、Kulldorffの方法は過疎地での集積性の検出力が大きかった。Kulldorffの方法は、地域の広がりの程度を評価するため、中には危険度が低い市町村まで高危険度と判定される場合もあり、集積地域の推定に問題があるが、広域的な施策体系の企画・立案には大変有意義と考えた。一方、Tangoによる方法は、高危険度の中心点を検定しているため理解が得られやすく、種々の調査や行政施策の費用対効果もあがりやすい。また、高死亡率地域のみならず低死亡率地域まで検出しているので、地域の疾病状況の全体像が把握できる点で大きなメリットがある点などが再確認できた。

A. 研究目的

今日公表され利用されている厚生統計指標は、年齢調整死亡率、標準化死亡比などのように市区町村などの地域の「人口の年齢分布の違い」を調整しているものの、「人口の大きさ」までは調整できていない。そのため、これらの指標を利用して数区分に色分けした疾病地図を作成すると、人口の小さい地域の指標のバラツキが大きく、わずかな死亡数の変化が見かけ上の指標を大きく変化させるという不安定性が指摘されている。本研究の第１の目的は、この問題を解決するために、近年、統計学の世界で計算機の発展によってその重要性が再認識され方法論が大きく進歩したベイズ流アプローチを利用して「人口の大きさ」を調整し、より適切な人口動態統計指標を開発しその利用を推進させることを目的とする。本研究の第２の目的は、疾病の地域集積度指標の開発である。限りある予算、資源を効率的に投入して地域ニーズに対応したきめ細かい対策の立案・実施を行うためには、対策が最も必要とされている最優先地区を選定する必要がある。この目的のためには、地域別に推定された疾病指標（ベイズ推定も含めて）では、大小に並べれば「必ず」最も高い地域が検出されるという意味で不適切であり、「疾病の集積性」を表現する別の指標を導入しなければならない。第３の目的は、これらの新しい指標の普及に向けた市区町村別疾病地図の視覚的表示解析システムを Windows 上で開発することにある。

B. 研究方法

今年度は，疾病地図における地域集積性を検出する二つの方法（Tango の検定とKulldorff の検定）について様々な集積モデル，地域による違いを検討するために，(1) New York, Boston, Washington などの都市を中心としたアメリカ東部地域にいて疾病集積モデルに基づくシミュレーションによる比較をKulldorff との共同研究で，また，(2) 日本の平成５年から９年までの人口動態統計死亡データ、特に，高知県における死亡状況を解析するなかで比較検討する。アメリカ東部については，図１，図２に示すような，Hot spot cluster（図１），Global chain cluster model

（図２）を仮定したシミュレーションモデルで比較検討を行った。疾病の集積モデルとしては、ある地域を中心とした集積モデル(hot spot cluster) と幾つかの地域の連鎖における集積モデル(chain cluster)、人口の大きさを考慮した地域モデルとして過疎地域、中都市、大都市を設定して比較をおこなった。その方法の詳細は別添の英文論文を参照されたい。

高知県の方法については，以下の通りである。

K-1) 調査地区及び時期

高知県内の５３全市町村における平成５年から平成９年までの死亡者（日本における日本人のみ）を対象とした。

K-2) 評価対象とした死因

あ）平成５年～６年

a. 胃の悪性新生物(I151)
b. 結腸、直腸、肛門の悪性新生物(I53–I54)
c. 肝及び肝内胆管の悪性新生物(I55)
d. 気管、気管支及び肺の悪性新生物(I62)
e. 虚血性心疾患(410-414)
f. 脳血管疾患(430-438)
g. 喘息(493)
h. 先天性代謝異常(740-759)

い）平成７年～９年

i. 胃の悪性新生物(C16)
j. 結腸、直腸、肛門の悪性新生物(C18-C21)
k. 肝及び肝内胆管の悪性新生物(C22)
l. 気管、気管支及び肺の悪性新生物(C33-C34)
m. 虚血性心疾患(I20-I25)
n. 脳血管疾患(I60-I69)
o. 喘息(J45-J46)
p. 先天性代謝異常(Q00-Q99)

(　)内の数字は、厚生省大臣官房統計情報部人口動態調査（死亡票））に使用されている原死因３桁基本分類コードを表す。

K-3) 評価方法

平成５年から平成９年までの５年間の人口動態調査（死亡票）を基に、新たに開発した市区町村別疾病地図の視覚的表示解析システム（Disease Mapping System）を使用し、SMR、SMRの経験的ベイズ推定量（$EBSMR$）、Tangoによる集積性の検定及び、Kulldorffによる集積性の検定を行った。

C. 研究結果

まず，アメリカ東部におけるシミュレーションの結果は別添英文報告に詳しく述べたが，Tangoによる方法と、Kulldorffの性質の違いがよく結果に現れた。結果は、Tangoの方法では大都市での集積性、chain cluster に対する検出力が高く、Kulldorffの方法は過疎地域，中都市での検出力が高い結果であった。ただ，Kulldorffの方法は集積していない周辺地域も巻き込んだ形での集積地域の推定を行ってしまう危険性が指摘できる。

一方，高知県における各死因ごとの総死亡数、期待死亡数、SMR、$EBSMR$の計算結果、SMRのχ^2検定結果及びTangoの集積性の検定結果、Kulldorffの集積性の検定結果は表１のとおりである。

1. 人口動態統計指標の評価（EBSMRとSMR）

図３に、胃の悪性新生物（男）を例に人口規模（観測死亡数）とSMR、$EBSMR$の計算結果を示した。これによると、人口規模が小さいほどSMRのばらつきが大きいが、$EBSMR$はその影響がなくなっている。表１の先天性代謝異常にみられるように、死亡例がなかった市町村の$EBSMR$は地域の平均値(100) に近づいている。特に高知県のように、人口規模が６００人から３０万人の市町村まで大きく開きがあって、小さい市町村では観測死亡数がごく僅かな疾病もみられ、市区町村単位で評価する場合は、SMRよりむしろ$EBSMR$によって評価するほうが妥当性はより高いと思われる。

2. 集積性の検定方法の評価（Tangoによる方法とKulldorffによる方法)

両者とも検出力には大きな違いはなかったが、図４，図５のように地域の広がりには差が出ている。Tangoの方法では、地域の中で最も相対危険度が高い地域を検出し、これに対しKulldorffの方法は相対危険度が高い地域の広がりを検出し

ている。その結果、両者の検定結果に差が出ている。また、Tangoの方法では、図５のように低死亡率側に有意な地域も検出している。

3. Tangoによる方法とKulldorffによる方法の疾病地図への適用

図６のように、一般的にはSMRを階級別に色分けして疾病地図を作製している例が多いが、この場合、ある特定地域の相対的な位置を主観的・視覚的に見分けはつくが、疾病集積性の客観的な判断材料(EBM: Evidence Based on Medicine)には応用できない。SMRのχ^2検定結果はその性質上、人口規模が大きいところが有意となる傾向が強いため、誤った認識を与える可能性がある。また、人口規模を調整した$EBSMR$の場合は、客観的な判断材料にはなるがそれでも情報量が多すぎる。その点、Tangoによる方法とKulldorffによる方法では、表現方法にこそ違いはあるものの、情報をよく集約できており、集積性のある地域を客観的、視覚的によく表現できている。

D. 考察

これまでも、SMRは人口規模による影響が大きく、人口動態評価指標としては$EBSMR$が優れた指標であることはよく指摘されていた。しかし、その計算方法はまだまだ一般的でなく、集積性の検定方法も含め、汎用的なアプリケーションの開発が強く望まれていた。今回、SMR、$EBSMR$、Tangoの集積性の検定、Kulldorffの集積性の検定を総合的に計算できる汎用的なアプリケーション（Disease Mapping System）の開発ができたこと、さらに、その実用性が証明できたことにより、死亡指標の客観的な評価手法は一歩前進したと考える。

今後は、このような科学的な判断材料(EBM)に基づく疾病対策の展開が望まれるところである。

E. 結論

1）本研究で検討した人口動態統計死亡指標の経験ベイズ推定、集積性の検定はこれまでの指標では検討できない優れた特徴があることを人口動態統計データ，シミュレーションなどにより検証した。

2）集積性の評価手法として用いた Tangoによる方法と、Kulldorffの性質の違いがよく結果に現れた。Tangoの方法は大都市の集積性に，過疎地域での集積性の検出には Kulldorffの方法がより適していることが示された。その利用に当たっては，その目的・性質をよく理解して適用することが重要である。

3）また，Tango の方法は、高危険度の中心点を検定しているため理解が得られやすく、種々の調査や行政施策の費用対効果もあがりやすく、かつ高死亡率地域のみならず低死亡率地域まで検出しているので、地域の疾病状況の全体像が把握できる点で大きなメリットがあると考えられる。

F. 研究発表

1)Tango, T. Extended score tests for focused clustering. The XXth International Biometric Conference, July, Berkeley, California, U.S.A. 2000.7 ; p93

2)中谷実，丹後俊郎 青森県における疾病の地域集積性. 第59回日本公衆衛生学会，群馬，2000.10, p829.

Power Comparisons for Disease Clustering Tests

Martin Kulldorff*, Toshiro Tango†

Abstract

Many different methods have been proposed to test for geographical disease clustering, and more generally, for spatial clustering of any type of observations while adjusting for an inhomogeneous background population generating the observations. Despite the many proposed test statistics, there has been few formal comparisons conducted. We present a collection of 1,220,000 simulated benchmark data sets generated under 51 different cluster models and the null hypothesis, to be used for power evaluations. We then use these data sets to compare the power of the Spatial Scan Statistic, the Maximized Excess Events Test and Bonetti-Pagano's M. All have good power, the former having an advantage for localized hot spot type clusters and the two latter for global clustering where randomly located cases generate other cases close by. By making the simulated data sets publicly available, new tests can easily be compared with previously evaluated tests by analyzing the same benchmark data.

Key Words: Spatial Statistics, Power, Geography, Spatial Epidemiology, Hypothesis Testing, Cluster Detection

*Division of Biostatistics, Department of Community Medicine and Health Care, University of Connecticut School of Medicine, 263 Farmington Avenue, Farmington, Connecticut 06030-6325, USA. Tel: 860-679-5473, Email: martink@neuron.uchc.edu

†Division of Theoretical Epidemiology, Department of Epidemiology, The Institute of Public Health, 6-1 Shirokanedai 4 chome, Minato-ku, Tokyo 108, Japan. Tel: 03-3441-7111, Email: tango@iph.go.jp

1 Introduction

A large number of different tests for spatial randomness that adjust for an uneven background population have been proposed. Such test statistics are, among other things, used to test whether the geographical distribution of disease is random or not, adjusting for the geographical distribution of the population at large. They are also used in areas such as archaeology, botany, criminology, demography, ecology, economics, engineering, forestry, genetics, geography, history, neurology, sociology and zoology. Several review articles have been written [4, 13, 26, 16, 28, 30, 32, 35, 41], the most extensive having identified over 100 different test statistics [27].

There has been few systematic comparative evaluations of tests for spatial randomness. Different tests have sometimes been applied to the same data sets [1, 11, 14, 39, 48, 53], but for a formal comparison of test statistics it is important to evaluate their power [52], and only a small fraction of the proposed tests have undergone such evaluations. Three major considerations when designing a power comparison study are (i) the reproducibility of the clustering process, (ii) the clustering models considered as the alternative hypotheses, and (iii) minimization of the bias and variance when estimating the difference in power for different tests.

1.1 Reproducibility

While very important, simulated power comparisons are tedious, time consuming and unglamorous to perform. Each of the methods to be evaluated must be programmed, the simulated data must be gener-

ated, and each test statistic must be calculated for each simulated data set. If there are previously published power evaluations, it is sometimes possible to avoid redoing the calculations for already evaluated test statistics, but that requires that the earlier simulation models are described in complete detail, which is seldomly the case. The ideal though is to go one step further, and build on previous power evaluations using not only the same alternative models but also the exact same simulated data. That minimizes the random variation as the methods are compared.

In this paper, we present and provide access to a set of benchmark simulated data sets. Using this benchmark, we evaluate the power of two test statistics. Other researchers can then easily compare tests of their interest to previously evaluated test statistics, by simply reanalyzing the benchmark data sets. This is the most economical way to conduct power comparisons of many different test statistics. Past evaluations of tests spatial randomness have for natural reasons been done mostly as pairwise power comparisons or more rarely in groups of three or four [21, 34, 38, 42, 43, 44, 45, 49]. By establishing the benchmark data sets, any new test evaluated will automatically be compared with all previously evaluated test statistics.

1.2 Clustering Models

With one exception, earlier power comparisons all considered first-order clustering models where cases are located independently of each other, but where the relative risk is different in different geographical areas. Most of these evaluated the power for a clustering model with one [21, 38, 42, 43, 44, 45], two [42, 43, 45], three [45] or four [42] hot-spot clusters; while Oden [34] used a clustering model with a different relative risk in each census area. As a contrast, Vach [49] considered a second-order clustering model where the location of one case is dependent on the location of other previously generated cases, at the same time as the risk varies geographically. There has not been any power comparison using a pure second-order model, where each particular case is randomly located, so that the relative risk is constant throughout the map, but where the location of cases are dependent on each other. It is important to realize that while first and second order models are very different in how the points are generated, the resulting point patterns may be exactly the same, and hence indistinguishable. Bailey and Gatrell (1995, chapter 3) provide an excellent discussion of this.

In this study we use 51 different clustering models, 15 with a single hot-spot cluster, 10 with multiple hot-spot clusters, and 26 with purely second-order clustering models where the risk is constant throughout so that any one particular case is spatially randomly located, but where the location of different cases are dependent on each other. For each model, the power is calculated conditioned on two different levels of the total number of cases. The number of alternative clustering models considered have in past studies been in the range of 3 to 8, with the exception of Vach [49] and Rogerson [38], who considered 12 and 20 different clustering models respectively. Except for the study by Tango [43], the power was never evaluated for different number of cases within the same model.

Another important aspect of a clustering model is the background population used. We use a real data set consisting of all women in 245 counties in Northeastern United States during 1988-1992. This is a fairly typical epidemiological data set, with data aggregated into a mix of rural and urban census areas.

1.3 Minimization of Bias and Variability

For some tests it is possible to evaluate the power using an asymptotic approximation of the test statistic distribution [34, 38, 43]. Unfortunately, asymptotic approximations do not exist for most test statistics. When they do exist, the asymptotics may be defined in terms of the geographical area, the population size or the number of cases going to infinity, with the other two parameters held at a specific constant or rate, and the approximations must be interpreted considering these asymptotic concepts. Unless the approximations for all test statistics are very good, it is necessary for comparison purposes to obtain the critical values through a large number of simulated data sets randomized under the null hypothesis. In

this paper we present two groups of 100,000 simulated data sets to estimate the critical values, with 600 and 6000 cases respectively.

In order to minimize the variability of the estimated power difference between tests, it is important to condition the analysis on a particular population distribution, and on the total number of cases. Moreover, different tests should be evaluated using the same random data sets.

Another factor determining the variance of an estimated power difference is the number of random data sets generated under each alternative hypothesis. As part of the benchmark data, we present 10,000 random data sets for each alternative.

1.4 Test Statistics Compared

Tests for spatial randomness can be classified based on their purpose. Focused tests are designed to test whether a local cluster exist around a predetermined point source, while general tests looks for clusters without any preconceived assumptions about their location [3]. Among general tests, cluster detection tests are used both to detect local clusters, without any preconceived idea of their location, and to determine their statistical significance. Global clustering tests, on the other hand, are used to determine whether there is clustering present throughout the study area, without determining statistical significance of individual clusters [26, 44].

Discussions regarding the differences between the latter two types of general tests have been provided elsewhere [26, 44], but their important difference is not always considered, and there has never been a formal study showing how they differ in terms of their power to detect different types of clustering. In fact, the power of global clustering tests have typically been evaluated using hot-spot cluster models. In this paper we evaluate the power of the Spatial Scan Statistic [24], the Maximized Excess Events Test [45] and Bonetti-Pagano's M [6]. These were chosen so as to not only compare three different tests, but equally important, to illustrate the differences between the two types of tests. We show that the Spatial Scan Statistic, a cluster detection test, has good power for hot spot cluster alternatives, while the Maximized Excess Events Test and Bonetti-Pagano's M, global clustering tests, have good power when clustering occurs throughout the geographical region of study.

Most tests for spatial clustering depend on a parameter that determines the size of clustering tested for. This includes the λ in Tango's Excess Events test [43], the k in Cuzick-Edward's k-nearest neighbor test [9], and the radius of the circle in Turnbull's CEPP [48]. The three tests compared in this paper do not depend on such a pre-specified parameter. This was the second reason for evaluating these particular test statistics. We expect that more such tests will be proposed as extensions of earlier methods, and it will then be of special interest to compare them with the tests evaluated here.

2 Benchmark Data Sets

For the benchmark data sets we use the female population in the 245 counties and county equivalents in the Northeastern United States, consisting of the states of Maine, New Hampshire, Vermont, Massachusetts, Rhode Island, Connecticut, New York, New Jersey, Pennsylvania, Delaware and Maryland, as well as the District of Columbia. Each county is geographically represented by a centroid coordinate. As the population for each county we used the number of women living there according to the 1990 United States census. This data has previously been used to evaluate the existence of geographical clusters of breast cancer mortality [22]. Both the population data and the geographical coordinates are available at 'http://www.commed.uchc.edu/biostat/datasets.htm'.

2.1 Hot Spot Clusters

For the hot spot alternatives, we constructed three different sets of local clusters in a rural, urban and mixed area respectively. Within each of these three sets, we constructed five different size clusters using 1, 2, 4, 8 and 16 counties. The center of the rural cluster was Grand Isle County in northern Vermont, on the Canadian border. Among the 245 counties, Grand Isle has the smallest population. The center of

the mixed cluster was Pittsburgh (Allegheny County) in western Pennsylvania. Pittsburgh is a large city, surrounded by rural areas. The center of the urban cluster was Manhattan (New York County) in New York City, closely surrounded by other very urban counties. Additional counties were added to the central county by order of geographic distance between county centroids. The clusters with 16 counties are shown on the map in figure 1. The New York City cluster is close to the population center of the region, while the Pittsburgh cluster contains the urban area furthest away from the population center.

The counties within each cluster were given a higher risk than the remaining clusters in the region. For each of the 15 clusters, the relative risks and the expected number of cases under both the null and the alternative hypotheses are given in table 1. The relative risks were set so that the null hypothesis would be rejected with probability 0.999 when using a standard binomial test, had we known the 'cluster counties' apriori, not taking the multiple testing into account. Let n be the combined population in the cluster counties, and let N be the total population in all counties. Conditioned on the total number of cases C, the observed number of cases in the 'cluster counties' is under the null hypothesis binomially distributed with mean $m_0 = Cn/N$ and variance $v_0 = C\frac{n}{N}\frac{N-n}{N}$. Using the normal approximation for the binomial distribution, the critical number of cases k needed in order for a one-sided test to reject the null hypothesis at the 0.05 level is then the k such that $\frac{k-m_0}{\sqrt{v_0}} = 1.645$. Under the alternative hypothesis with a relative risk of r for the 'cluster counties', the number of cases in those counties is binomially distributed with mean $m_A = C\frac{nr}{N-n+nr}$ and variance $v_A = C\frac{nr}{N-n+nr}\frac{N-n}{N-n+nr}$. Using the normal approximation again, we then selected the relative risk r such that $\frac{k-m_A}{\sqrt{v_A}} = 3.09$. This choice of relative risks provides an upper limit of 0.999 for the power attainable by any test for spatial clustering, and a yard stick by which to compare the performance of a test statistic on different hot-spot clusters.

In order to evaluate how the disease clustering tests perform when there are multiple hot-spot clusters, we constructed five alternative models that included one rural and one urban cluster, with the same number of counties in both clusters. For another five alternative models, we also included the mixed cluster with an identical number of counties. For each cluster we set the relative risks as before, according to Table 1. Note that while the number of counties are the same in each cluster, the relative risks and the population sizes are different.

In total we constructed 25 different hot-spot cluster models, with varying characteristics.

2.2 Global Chain Clustering

For the global clustering alternative, we want cases to be clustered wherever they occur in the region. Moreover, for all counties we want the expected number of cases to be the same under the null and alternative hypotheses. These requirements pose a special challenge in constructing a clustering model.

For the global clustering model a certain number of cases are first located randomly on the map, according to the null hypothesis. These original cases then generate other new cases close by. If each original case generates one additional case, we call them twins, and if two additional cases are generated, we call them triplets.

Let n_i be the population of county i, and let $N = \sum_i n_i$. Let d_{ij} be the Euclidean distance between counties i and j. If the original case is in county i, a natural way is to assign its twin is to put it in county j, where j is chosen so that $\sum_k I(d_{ik} < d_{ij})n_k < rN \leq \sum_k I(d_{ik} \leq d_{ij})n_k$, for some constant $0 \leq r < 0.5$. This means that the twin is selected as the rN-nearest neighbor of the original case. In other words, a randomly selected case has probability r of being closer to the original case than what the twin is. A problem with this type of approach is that the additional cases will not be spatially randomly distributed, but have a higher chance to occur in the central areas of the map as compared to outlying areas. This is because someone in the center of the map is a closer neighbor to more other individuals as compared to someone that lives close to the border. Hence, the requirement that every county has the same expected number of cases under the null and alternative hypotheses is not met.

To overcome the above problem, we used what we call a global chain clustering model. The counties are tied together sequentially on a chain that passes through each county exactly once, after which it reconnects with the first county on the chain, forming a Hamiltonian cycle. The randomization of twins and triplets is then embedded within this chain, so that an additional case is assigned to county j if $\sum_k I(d'_{ik} < d'_{ij})n_k < rN \leq \sum_k I(d'_{ik} \leq d'_{ij})n_k$, where d'_{ij} is now the distance in one particular direction along the chain connecting the counties. hence, the twins are assigned as the rN nearest neighbor along one direction of the chain. For twins, the probability model is the same independent of the direction used. For triplets, the two new cases were assigned in opposite directions. Note that the chain does not imply that the 'disease' spreads itself around the chain, just that twin and triplet cases are located in either of only two directions, as defined by the chain.

The chain used is shown in figure 2. The chain was constructed so that two counties next to each other on the chain always border each other geographically. Moreover, it was constructed so that all counties in a state occur consecutively along the chain, except for New York and New Hampshire where that was not possible as these two states stretch from the Canadian border to the Atlantic coast. Within these parameters, the exact construction of the chain is arbitrary, but an attempt was made to have counties that are geographically close to be close on the chain as well, to the largest extent possible.

We constructed different clustering models by using different constants r for the population based distances between the twins. In the most clustered scenario, the distance was zero, so that twins are always assigned to the same county. The chain was not needed for this scenario. In a second set of clustering models, r was set to be fixed at 0.005, 0.01, 0.02, 0.04, 0.08 and 0.16 respectively. In a third set of clustering models, r was exponentially distributed with mean 0.005, 0.01, 0.02, 0.04, 0.08 and 0.16 respectively.

With the chain model, to use $r = 1$ is the same as using $r = 0$. To use $r = 0.5$ would assign the twins at the opposite ends of the chain, typically putting them further away from each other than they would be by chance, leading to the opposite of clustering. If the chain was a perfect circle with even population distribution along the circle, then $r = 0.22$ would correspond to a situation where the expected distance between twins is the same as the expected distance between any two cases under the null model (see appendix). If the chain was circular, with an uneven population distribution, that equality would hold for a smaller r. In our case, we don't have a circular chain so it is not clear what value of r represents a situation of no clustering, but the above reasoning for a perfectly circular chain means that we should not necessarily expect to see clustering for distances greater than $r = 0.22$, and the largest distance used at $r = 0.16$ represents at most a very weak amount of clustering.

The equivalent 13 models were also used for triplets, resulting in a total of 26 global chain clustering models.

2.3 Simulated Data

In order to perform power comparisons, we constructed a number of random data sets. These are in two groups, with 600 and 6000 simulated cases respectively. These numbers were chosen, because we wanted the total number of cases to be divisible by both 2 and 3, to fit with both the twin and triplet models. The same data sets were used to evaluate power both at the $\alpha = 0.05$ and $\alpha = 0.01$ significance levels.

The same null hypothesis is used throughout, where the relative risk is set to one for each county, and case locations are independent of each other. This means that a particular case is assigned to county i with probability n_i/N. We generated 100,000 random data sets with 600 cases and the same number of data sets with 6,000 cases. These are used to estimate the cut-off point for significance. For each alternative hypothesis, we generated 10,000 random data sets. Using the null cut-off points, these were used to estimate the power. A Lehmer random number generator was employed, with modulus 2,147,483,647 and multiplier 48,271 [37].

The same data sets were used for the three tests, so as to eliminate any power differential

due to some data sets being by chance more clustered than others. All 1,220,000 data sets can be downloaded from the world wide web at 'http://www.commed.uchc.edu/biostat/datasets.htm'.

3 A Cluster Detection Test and a Two Global Clustering Tests

The clustering models described above can be used for power analyses for any number of disease clustering tests. In this paper we estimate the power of the Spatial Scan Statistic [24], the Maximized Excess Events Test [45] and Bonetti-Pagano's M [6]. As others use the same data sets to evaluate other disease clustering tests, they only need to do the power calculations for the new tests, enabling an automatic comparison with these three test statistics.

3.1 The Spatial Scan Statistic

The Spatial Scan Statistic [24] imposes a circular window on a map and lets the circle centroid move across the study region. For any given position of the centroid, the radius of the window is changed continuously to take any value between zero and some upper limit. In total, the method uses a set $\mathcal{Z}$ containing an infinite number of distinct circles, each with a different location and size, and each being a potential cluster. We set the upper limit so that the circle may contain at most 50 percent of the total population.

Under the alternative hypothesis, there is at least one circle for which the underlying risk is higher inside the circle as compared to outside. For each circle, it is possible to calculate the likelihood to observe the observed number of cases within and outside the circle respectively. The circle with the maximum likelihood is defined as the *the most likely cluster*. This is the cluster that is least likely to have occured by chance.

The likelihood can be calculated assuming either a Poisson or Bernoulli model, depending on how the cases are generated. We use the former.

Let $c(Z)$ be the observed number of cases in circle Z. Let $n(Z)$ be the expected number of cases in circle Z under the null hypothesis, so that $n(A) = c(A) = C$, where A is the total region under study. Let $L(Z)$ be the likelihood under the alternative hypothesis that there is a cluster in circle Z, and let L_0 be the likelihood under the null hypothesis. It can then be shown that

$$\frac{L(Z)}{L_0} = \left(\frac{c(Z)}{n(Z)}\right)^{c(Z)} \left(\frac{C - c(Z)}{C - n(Z)}\right)^{C - c(Z)} \quad (1)$$

if $c(Z) > n(Z)$ and one otherwise. Details, including derivations as a likelihood ratio test, have been given elsewhere [24]. As this likelihood ratio is maximized over all the circles, it identifies the one that constitutes the most likely cluster. The test statistic is

$$\max_Z \frac{L(Z)}{L_0}.$$

When derived as a likelihood ratio test, it is based on a set of alternative hypotheses, each with a single circular cluster of different size, location and relative risk. This does not mean that the test statistic can only detect circular clusters, but should expect higher power for more compact clusters if everything else is equal. Its p-value is obtained through Monte Carlo hypothesis testing [12]. Calculations were done using SaTScan [25]. The method has been applied in a variety of epidemiological studies [17, 18, 19, 20, 22, 33, 40, 50, 51].

3.2 The Maximized Excess Events Test

Let c_i be the observed number of cases in county i, and let $C = \sum_i c_i$ be the total number of cases. Let n_i be the expected number of cases in county i under the null hypothesis, so that $\sum_i n_i = C$. For a given constant λ, the Excess Events Test statistic [43] is defined as

$$EET(\lambda) = \sum_i \sum_j a_{ij}(d_{ij}, \lambda)(c_i - n_i)(c_j - n_j)$$

where

$$a_{ij}(d_{ij}, \lambda) = e^{-4d_{ij}^2/\lambda^2}$$

and d_{ij} is the Euclidean distance between counties i and j. Other choices of $a_{ij}(d_{ij}, \lambda)$ are also possible [34, 43, 38]. The choice of λ relates to the geographical scale of clustering, and is to some extent arbitrary. A large λ will give a test sensitive to geographically large clusters, while a small λ will make the test more sensitive to small ones.

To be able to detect clustering irrespectively of its geographical scale, Tango [45] proposed the Maximized Excess Events Test (MEET):

$$MEET = \min_{0 \leq \lambda \leq U} P[EET(\lambda) > eet(\lambda) | H_0, \lambda]$$

where $eet(\lambda)$ is the observed value of the Excess Events Test statistic conditioning on λ, and U is an upper limit on λ. Practical implementation of the test uses 'line search' by discretization of λ, and the MEET statistic is evaluated using Monte Carlo hypothesis testing [12].

Calculations were done using a specially written S-Plus code [46]. The method has been applied to various epidemiological data sets [18, 20, 33, 47].

3.3 Bonetti-Pagano's M

3.4 Clustering Concepts

The three test statistics uses somewhat different clustering concepts as part of their definitions. The spatial scan statistic evaluates the strengths of individual circular clusters, but without any assumptions concerning their size or location. Only the cluster of maximum strength is considered when evaluating the null hypothesis. This means that if the null-hypothesis is rejected, it is possible to pinpoint the location of the cluster causing the rejection, as it is independent of clusters occuring in other regions of the map [24]. The MEET statistic evaluates whether excess of cases occur in close geographical proximity to each other. Evidence of clustering is summed over the map as a whole, so that multiple small clusters may together cause a rejection of the null hypothesis, even if they are far from each other. Bonetti-Pagano's M ...

4 Power Comparison Results

4.1 Hot Spot Clusters

The results of the power analyses for the hot-spot clusters are shown in Table 2. For the rural clusters, the Spatial Scan Statistic has very high power while the power of the Maximized Excess Events Test was low. For the mixed clusters, both tests have high power with a slight advantage for the Spatial Scan Statistic. Both tests also have high power for the urban clusters, now with a slight advantage for the Maximized Excess Events Test.

Both tests have higher power when there are two or three different hot spot clusters in the same model. This is expected since more clustering is introduced when additional clusters are added to a model. The Spatial Scan Statistic has, with some exceptions, slightly higher power, but the difference is small.

Comparing the power between different cluster models, we find that the Spatial Scan Statistic has highest power for the rural cluster models, followed by the mixed, and then the urban. The Maximized Excess Events Test on the other hand, has highest power for the urban cluster models, followed by the mixed and the rural. Does this mean that the power is a function of a cluster's population size? Not necessarily. Within the mixed cluster models, the opposite relationship occurs. The power of the Spatial Scan Statistic increases with increasing population size, while for the Maximized Excess Events Test, the power sometimes decreases with increasing cluster population size. The power is not a simple function of a cluster's population size, but is also related to the geographical size of the cluster, as well as to the level of spatial aggregation (i.e. the average county population size) in and around the cluster.

4.2 Global Chain Clustering

The power estimates for global chain clustering are shown in Table 3. When the distance is zero, both tests have good power but with a clear advantage for the Maximized Excess Events Test. As expected, the power goes down with increasing distance between twins and between triplets, and approaches the

nominal significance level at the greatest distance of $r = 0.16$. Note that the power is consistently higher for the clustering models in which the distances between twins/triplets are random according to the exponential distribution, as compared to the fixed distance model, even though the expected distances are the same.

5 Discussion

The results of this study clearly show how different disease clustering tests are good for different types of alternative hypotheses. If one is interested in detecting and evaluating localized clusters, it is better to use the Spatial Scan Statistic, while the Maximized Excess Events Test is better at detecting global type clustering that is present throughout the study region. This is to some extent intuitive.

The Maximized Excess Events Test is based on the evidence of clustering found throughout the map, as the test statistic is a summation over all the counties. When a cluster is large in population size though, it also performs well for hot spot clusters, since there is then a large proportion of the population that is affected by the cluster. The test statistic uses geographical distance to define closeness of cases. This may explain why it has higher power for hot-spot clusters in urban as opposed to rural areas, as the latter clusters are more dispersed. It also means that it may perform better for a global chain clustering model where distance between twins are defined in terms of geographical distance rather than the nearest neighbors. Although not shown here, a feature of the Maximized Excess Events Test that is of great additional value is that it is possible to determine how much each county contributes towards the total amount of clustering observed, as represented by the magnitude of the test statistic corresponding to that particular county [45].

The Spatial Scan Statistic ignores all the information about the location of cases except whether the case is inside or outside the currently evaluated circular zone. The disadvantage of this is that the power is lower when clustering occurs throughout the study region. An advantage is that the rejection can be wholly attributed to a particular cluster, since any rearrangement of the cases outside the cluster cannot reduce the value of the test statistic, no matter how the rearrangement is done. This is discussed elsewhere in formal mathematical language [24]. By use of circles, the power depends on the compactness of the cluster shape. The true cluster need not be circular to obtain good power, but the test should not be expected to have good power for a long and narrow cluster, such as along the Hudson river.

Hence, the two tests put weight on different aspects of clustering, and we can classify the Spatial Scan Statistic primarily as a cluster detection test, and the Maximized Excess Events Test as primarily a global clustering test.

After rejecting the null hypothesis, concluding that there is some form of clustering, it is of course of interest to know the exact nature of the clustering process. For example, is it global type clustering or are there hot-spot clusters? If the former, do the cases consist of twins, or triplets, or do they consist of small groups with a variable number of cases, or are all cases generated through one single process where each new case generates another one? If the latter, how many hot-spots are there and where are they located? It is important to note that the power estimates provided reflect the power to reject the null hypothesis for whatever reason and that the probability of both rejecting the null hypothesis and correctly determining the type of clustering process is a different matter.

Other scientists are encouraged to use the benchmark data sets presented in this paper to evaluate disease clustering tests that they consider using, or to create new tests that will perform better than those evaluated here. Existing tests of potential interest include the k-Nearest Neighbors Test [7, 9], Swartz' Entropy Test [42], Besag-Newell's R [3], the Isotonic Spatial Scan Statistic [23], Grimson's Method [15], Martuzzi-Hills' Gamma Method [31], Oden's I_{pop} [34], Rogerson's R [38], Ord and Getis' max G_i [36], Diggle-Chetwynd's D [10] and Bithell's M [5]. It would also be worth investigating the Maximized Excess Events Test with other weight functions $a_{ij}(d_{ij}, \lambda)$. Comparisons are of great interest regardless of whether other tests turn out to have greater

or lower power than those presented here, as it will spread light on the question of what types of tests are good for what types of clustering models.

It is important to keep in mind that any simulated power comparison is dependent on the particular data set and alternative models used. Most tests will have relative strengths and weaknesses for different clustering models and a single test cannot have optimal power for all alternative hypotheses.

For this study we used a data set typical of epidemiological applications, where both the population and cases are aggregated into census areas of different population size. A limiting factor is that only one set of spatially distributed populations numbers was used, and the strength of various test statistics may not only depend on the alternative clustering models, but also on the spatial distribution of the aggregated areas as well as the relative population sizes in these areas. In terms of different number of cases, the comparative results were very similar for 600 and 6000 respectively, so the sensitivity to this model parameter is of lesser concern.

This paper can be viewed as presenting only a first batch of simulated benchmark data sets for disease clustering test evaluations. Others investigators are encouraged to contribute simulated data generated from other alternative hypotheses of interest. Ideally, this will produce a collection of simulated benchmark data sets for the communal use of all researchers in this area. Other clustering models to consider may be (i) interior hot-spot clusters, (ii) hot-spot clusters with different levels of risk in the center and peripheral areas, (iii) a long and narrow hot-spot cluster, (iv) a very large number of geographically small hot spot clusters, say about one or two dozen, (v) a global clustering model where each original case has a random number of 'siblings' rather than the fixed number that we used, and (vi) a global double-chain clustering model, with two separate disconnected chains covering two different parts of the map, such as the more rural and urban areas respectively, and with the strength of clustering being different within the two chains. One could also use a Cox Process [29, 8], where cluster locations and relative risks are random rather than deterministic. The advantage of this is that the comparison of the test statistics would reflect the average performance for a large group of different hot-spot clusters. The disadvantage is that one will not learn for what specific types of hot-spot clusters a particular test statistic has high or low power.

It would also be worth while to create benchmark data sets for non-aggregated data sets, where each case has unique coordinates.

There are many more tests for spatial clustering, and many more clustering models, for which it is worth while to estimate power. We hope that other researchers will build upon this work, and evaluate other tests using the clustering models used here and, equally important, that they will generate and share simulated data from other important clustering models. Most importantly, with the existence of a set of benchmark data sets, each new power comparison does not need to start from scratch, but can build upon previously calculated power estimates for previously evaluated tests.

6 Appendix

Suppose we have a circle with radius one centered at $(0,0)$. The distance from $(1,0)$ to the point on the circle corresponding to x degrees is

$$\sqrt{(1-\cos x)^2+\sin^2 x} = \sqrt{2-2\cos x}.$$

The distance to a point 22 percent along the circle is $\sqrt{2-2\cos(2\pi 0.22)} = 1.27$. The expected distance from $(1,0)$ to a random point on the circle is

$$\int_0^{2\pi} \sqrt{2-2\cos x}\; dx = 1.27$$

Acknowledgments

The authors thank Marco Bonetti for advice concerning the implementation of Bonetti-Pagano's M, and two anonymous reviewers for valuable comments that improved the quality of the paper.

References

[1] Alexander FE, Boyle P, editors, Methods for investigating localized clustering of disease,

IARC scientific publication no. 135, International Agency for Research on Cancer, Lyon, 1996.

[2] Bailey TC, Gatrell AC, Interactive spatial data analysis, Longman Scientific & Technical, Harlow Essex, England, 1995

[3] Besag J, Newell J, The detection of clusters in rare diseases, *Journal of the Royal Statistical Society*, **A154**, 143-155 (1991).

[4] Biggeri A, Marchi M, Metodi di analisi spazio-temporale in campo epidemiologico: una rassegna, in Zani (ed), *Metodi statistici per le analisi territoriali*, Franco Angeli, Milan, 1993.

[5] Bithell JF, Disease mapping using the relative risk function estimated from areal data, In Lawson et al. (eds), *Disease mapping and risk assessment for public health*, Wiley, London, 1999.

[6] Bonetti M, Pagano M, A distance-based method for the detection of clustering, manuscript (2001).

[7] Cliff AD, Ord JK, *Spatial autocorrelation*, Pion, London, 1973.

[8] Cressie NAC, Statistics for spatial data, Wiley, New York, 1993.

[9] Cuzick J, Edwards R, Spatial clustering for inhomogeneous populations, *Journal of the Royal Statistical Society*, **B52**, 73-104 (1990).

[10] Diggle PJ, Chetwynd AD, Second-order analysis of spatial clustering for inhomogeneous populations, *Biometrics*, **47**, 1155-1163 (1991).

[11] Draper G, editor, The geographical epidemiology of childhood leukemia and non-Hodgkins lymphomas in Great Britain, 1966-83, *Studies on medical and population subjects no. 53*, HMSO, London, 1991.

[12] Dwass M, Modified randomization tests for nonparametric hypotheses, *Annals of Mathematical Statistics*, **28**, 181-187 (1957).

[13] Elliott P, Martuzzi M, Shaddick G, Spatial statistical methods in environmental epidemiology: a critique, *Statistical Methods in medical Research*, **4**, 137-159 (1995).

[14] Glaser SL, Spatial clustering of Hodgkin's disease in San Francisco Bay area, *American Journal of Epidemiology*, **132**, S167-177 (1990).

[15] Grimson RC, Rose RD, A versatile test for clustering and a proximity analysis of neurons, *Methods of Information in Medicine*, **30**, 299-303 (1991).

[16] Heywood JS, Spatial analysis of genetic variation in plant populations, *Annual Review of Ecology and Systematics*, **22**, 335-355 (1991).

[17] Hjalmars U, Kulldorff M, Gustafsson G, Nagarwalla N, Childhood leukemia in Sweden: Using GIS and a spatial scan statistic for cluster detection, *Statistics in Medicine*, **15**, 707-715 (1996).

[18] Imai J, Spatial disease clustering in Kochi prefecture in Japan - evaluation of disease indices and disease mapping, *NIPH Epidemiology and Biostatistics Research 1998*, National Institute of Public Health, Tokyo, 57-96, 1998.

[19] Kharrazi M, Pregnancy outcomes around the B.K.K. landfill, West Covina, California: An analysis by address, California Department of Health Services, 1998.

[20] Kojima M, Spatial clusters of diseases and nutritional factors in Shiga prefecture in Japan, 1987-1996'. *NIPH Epidemiology and Biostatistics Research 1999*, National Institute of Public Health, Tokyo, 73-120, 1999.

[21] Kulldorff M, Nagarwalla N, Spatial disease clusters: Detection and inference, *Statistics in Medicine*, **14**, 799-810 (1995).

[22] Kulldorff M, Feuer E, Miller B, Freedman L, Breast cancer in northeast United States: A geographic analysis, *American Journal of Epidemiology*, **146**, 161-170 (1997).