約情報を記述するための XML フォーマットであり，ニュースやイベントなどのタイムリーな情報を効率良く発信する仕組みを提供する．いくつかのニュースサイトでは，RSS に基づくのフレームワークを採用し，Web 上のユーザにニュース情報を発信している．情報提供者は，ニュースの要約の URL を提示し，要約を更新する．RSS リーダーは，RSS 情報のチェックを支援するためのツールであり，関心のある記事の URL を定期的にチェックすることで，ユーザは容易に興味あるサイトの要約や更新情報を確認することができる．

最近では，多くのユーザが自分のブログサイトで，日記や写真などの個人的な空間情報を投稿しているが，RSS に基づく情報配信をサポートしているブログサイトも多い．このように RSS は，公共的なニュースサイトだけでなく，個人サイトが，タイムリーかつリアルタイムの情報を発信することを可能としている．

本研究の目的は，個人の空間的かつ時間的な経験の管理，探索，マッピング，発信のためのツールを開発することである．こうしたツールを利用することで，ユーザは，いつでも，どこでも，自分自身の行動を記録し，個人データベースから情報を引き出すことができるであろう．本フレームワークにより，一種の “ユビキタスマッピング” [5,6] を実現できると考える．

3. フレームワークの設計

3.1 Spatial RSS: パーソナル空間情報の記述

本フレームワークでは，RSS (RDF Site Summary) に基づいて，実世界および Web 情報空間から収集したパーソナル空間情報を記述する．RSS に基づく記述は，Semantic Web とも密接に関係するが，XML の名前空間を定義することで，柔軟かつ拡張性を持った記述が可能となる．

図 1 は，本研究で提案する spatial RSS の概要と実世界および Web 情報空間からのパーソナル空間情報の記述の流れを表している．通常の RSS と同様に，次の要素を含み，それぞれの要素が XML のタグに対応している．

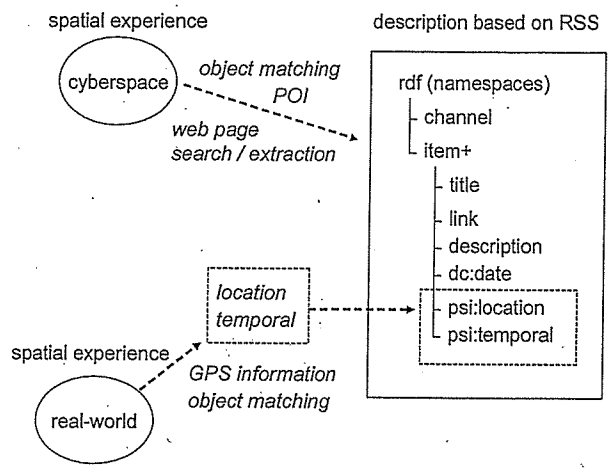


図 1　RSS に基づくパーソナル空間情報の記述

- channel：情報源に関する情報
- item：各空間オブジェクトに対応
- title：空間オブジェクトに関するタイトル情報
- link：空間オブジェクトに関連する URL
- description：空間オブジェクトの要約情報
- dc:date：オブジェクトに関する情報の更新時刻

これらの要素に加えて，次の 2 つの要素を含む．

- psi: location
- psi: temporal

psi:location は，空間オブジェクト位置を表現する要素であり，子要素として，id, name, point, line, area を持つ．name は空間オブジェクトの名前を表し，point は点座標を表す．Geo Vocabulary グループ[7]では，RSS ファイルの中での点情報の記述を提案しているが，本研究の spatial RSS は，線，面，名前による表現もサポートする．psi:temporal は，時間情報に関する要素であり，子要素に，time-stamp, period を持つ．time-stamp は，空間オブジェクトに関する時刻を表し，period は期間を表し，例えば，イベントの開始時刻や開催期間が記述される．現状では，基本的なタグ定義となっているが，今後，時空間情報を表現する様々な XML 仕様[7,8,9,10,11]を参考にしながら，拡張していきたいと考えている．

3.2 RSS の生成と POI 情報の登録

実空間における緯度経度を伴う空間経験は，その緯度経度とデータベース中の空間オブジェクトの位置情報の比較が行われ，もし該当する空間オブジェクトが存在すれば，そのオブジェクトに関連付けられ，更新情報が記述される．また，ユーザが実世界で認識した空間オブジェクトを新規の POI データとして登録することもできる．オブジェクトが既に登録されている場合でも，その位置情報が異なる場合には，点情報を追加していく．ユーザ自身が空間オブジェクトを，名前と緯度経度と共に登録していくことで，一種のユーザ辞書が構成されていくと考えられる．本フレームワークにおいては，探索やマッピングを行う前に，基盤となる POI データベースを構築することが重要である．

Web 情報空間での探索結果を空間経験として RSS 配信する場合，オブジェクトの名前により直接的にマッチングする方法と，アドレスマッチングなどのジオコーディングを適用する方法が考えられる．ジオコーディングにより緯度経度に変換する場合は，実空間における空間経験と同様の手順で，RSS ファイルを生成する．Web ドキュメントの場合でも，アドレスマッチングの手法を用いることで，POI データを登録することが可能である．この具体的な方法については，4.3 にて説明する．

3.3 Spatial RSS viewer

通常，RSS リーダーは，登録された Web ページについて，更新情報と要約情報をユーザに提供する．登録 Web ページのタイトルがリストとして表示され，ユーザが，リストの要素をクリックすることで，対応する Web ページが表示される．しかし，本研究では，RSS 記述が空間的な表現を含むため，地図上で空間オブジェクトに関する情報を表示できるツールが望まれる．本研究では，そのツールを，spatial RSS viewer と呼び，開発を進めている．

Spatial RSS viewer は，GPS 位置情報や空間オブジェクトの名前によるマッチングにより，実世界の空間的な経験を示し，空間オブジェクトとのマッチングによってウェブ探索の結果を表示する．そして，これらの空間オブジェクトについてのタイムリーな情報を地図上に表示する．もしユーザがデスクトップ端末上で，特定の領域中の空間オブジェクトを監視したい場合，あるいは，ナビゲーション中に移動ユーザからの周辺情報を探索したい場合には，spatial RSS viewer を利用することで更新情報を取得できると考える．

4. 実装および考察

4.1 プロトタイプシステムの構成

図 2 にプロトタイプシステムの構成を示す．

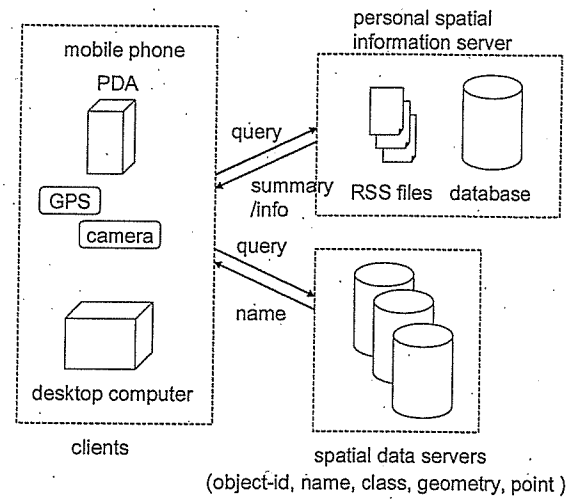


図 2　システム構成

プロトタイプシステムは，次の要素から構成される．

- モバイル・デスクトップクライアント

サーバに蓄積された実空間および Web 情報空間からの収集結果を閲覧するだけでなく，個人の空間的な経験あるいは POI データをサーバに登録する．

- パーソナル空間情報サーバ

ユーザからのパーソナルな空間情報をデータベースで管理し，空間経験を表す RSS ファイルを生成し，公開ディレクトリに配置する．

- 空間データサーバ及び POI データベース

クライアントが空間オブジェクトの名前を取得する際に，参照するサーバである．ユーザからの POI 登録によるデータベースだけでなく，国土数値情報や空間

データ基盤などの数値地図に基づくデータベースも想定する.

4.2 実空間における POI 情報の収集と登録

4.2.1 GPS ケータイによる POI 情報の登録

au の GPS 付き携帯電話の EZ ナビウォーク というアプリケーションには,「現在地メール」という機能がある. GPS 測位の結果である緯度・経度 (実際には, その緯度経度をパラメータとした地図サイトへのリンク) をメール送信することができる. 例えば,「件名」に空間オブジェクト名,「内容」に簡単な記述を加え, サーバ側での受信メールを解析することによっても, POI 情報の収集が可能である. ただし, 空間オブジェクトの名前に注目した場合, この方法では, 同一オブジェクトに対して, 表現の異なる多数の名前が登録されてしまう可能性がある.

本研究では, ユーザとの対話的なインタラクションを重視し, CGI を利用したアプローチを取る. au の GPS ケータイでは, 測位結果である緯度・経度を, CGI のパラメータとして渡すことが可能であるため, 現在, POI 登録および RSS 配信のための CGI を作成中である. 実験端末として, SANYO 製の W22SA を用いている. 以下, 動作の概略を示す.

1. GPS ケータイにおいて, 位置情報を測位.
2. 測位結果を CGI の引数としてサーバにアクセス.
3. サーバ側の CGI は, 受信した緯度・経度情報に基づいて, 候補となる「空間オブジェクト」の名前をメニューリストとして生成.
4. 携帯電話のブラウザ上で, 空間オブジェクトの候補となる名前がメニューとして表示.
5. 候補の中に, 該当するものあれば選択, もしなければ, 新規に入力する.
6. タイトルおよび説明欄を入力し, [登録]ボタンを押すと, サーバに情報が送信され, POI データが登録される.

本研究では, GPS ケータイを, 記事投稿のツールだけではなく,「空間オブジェクト」の名前を収集し, データベース化するツールの一つとして利用する. サーバ側で, ユーザの発信した位置情報とデータベース中の POI データの位置情報を比較し, 候補となる名前を提示することで表記の揺れをある程度抑制できると考えられる. また, 観測・興味の対象を,「空間オブジェクト」として識別できれば, 緯度経度情報を取得できなくても, そのオブジェクトに対する記事投稿が可能になると考えられる.

4.3 Web 情報空間からの POI 情報の収集と生成

4.3.1 空間ドキュメント管理システム

著者らは, 空間ドキュメント管理システム (SDMS: Spatial Data Management System) の設計と開発を進めている. SDMS は, 従来の構造が厳密な空間データではなく, 未構造や半構造の一般的なドキュメント (HTML, TEXT, EXCEL, PDF など) を対象とする. これらのドキュメントから住所情報を抽出し, アドレスマッチングにより緯度経度に変換して POI を作成する (図 3).

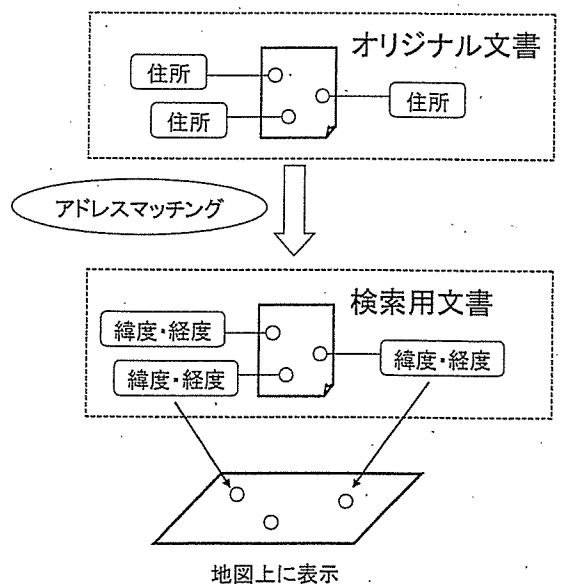


図 3 空間ドキュメントの POI 化

4.3.2 SDMS による POI 化

ユーザが, Web 検索の結果である HTML ページ (Web ドキュメント) を保存し, そのファイルを,

SDMS にドラッグ&ドロップすることで，対象ドキュメントがPOI化され，点データを地図上に表示することができる．Web ドキュメントに対して，緯度経度情報が付加されたため，キーワード検索だけでなく，空間範囲検索をすることも可能である．

SDMS はエキスポート機能も持っており，現状では，G-XML2.0 と Shape ファイル形式での出力が可能であるが，リモートサーバに対する出力機能を持たせることができれば，その情報に基づいて，RSS 配信することも可能であると考える．ただし，POI 登録の際は，空間オブジェクトの名前を抽出することが一つの課題となる．本フレームワークとの融合も考慮して，SDMS の開発を進めていきたいと考えている．

5. まとめ

本稿では，実空間と Web 情報空間から収集したパーソナル空間情報の記述・発信・閲覧のためのフレームワークを提案した．空間オブジェクトに関する情報は RSS に基づいて記述され，位置と時間に関する表現を含む．位置情報は点だけでなく，線や面データを記述し，空間オブジェクトの名前を探索やマッピングのために利用する．パーソナル空間情報の共有と交換のための RSS に基づくフレームワークと仕組みを開発し，ユビキタスマッピングのためのサポートツールとして広めたいと考えている．

謝辞

空間ドキュメント管理システムの研究開発は，厚生労働省科学研究費補助金（健康科学総合研究事業）「地理及び社会状況を加味した地域分析方法の開発に関する研究」（代表：浅見泰司）の支援を受けています．開発に際しては，東京大学生産技術研究所 相良 毅 助手，および，株式会社ジャスミンソフトからご協力頂いています．

参考文献

[1] 白石陽・安西祐一郎 (2004) パーソナル空間情報システムのためのセンサデータマッピングフレームワーク，「電子情報通信学会論文誌」, J87-A,1, 96-107.

[2] 伊藤昌毅・徳田英幸 (2003) ユーザの行動を反映した位置履歴表示システムの構築，「マルチメディア,分散,協調とモバイル（DICOMO）シンポジウム」.

[3] 杉本智彦 (2002)『山と風景を楽しむ地図ナビゲータ カシミール3D GPS応用編』，実業之日本社

[4] RDF Site Summary (RSS) 1.0, RSS-DEV Working Group, http://purl.org/rss/1.0/spec/

[5] Commission on Ubiquitous Mapping, International Cartographic Association, http://www.ubimap.net

[6]"International Joint Workshop on Ubiquitous, Pervasive and Internet Mapping (UPIMap)", Japan, Sep.2004

[7] W3C, Geo Vocabulary, http://w3c.org/2003/01/geo

[8]GeoOntologies, http://www.mindswap.org/2004/geo/geoOntologies.shtml

[9] ISO/TC211, Geographic information/Geomatics, http://www.isotc211.org/

[10] G-XML, http://gisclh.dpc.or.jp/gxml/

[11]GML, Open GeoSpatial Consortium, http://opengis.net/gml/

空間ドキュメント管理システムの設計と開発に関する研究

－SDMS（ Spatial Document Management System ）－

浅見 泰司　有川 正俊　白石 陽　片岡 裕介　相良 毅
東京大学 空間情報科学研究センター
東京大学 生産技術研究所

２００５
CSIS DAYS 2005
Research Abstracts On Spatial Information Science

空間ドキュメント管理システムの設計と開発に関する研究
- SDMS (Spatial Document Management System) -

浅見 泰司[1], 有川 正俊[1], 白石 陽[1], 片岡 裕介[1], 相良 毅[2]
[1]東京大学 空間情報科学研究センター, [2]東京大学 生産技術研究所
連絡先: <arikawa@csis.u-tokyo.ac.jp> Web: <http://www.s-it.org/sdms/>

(1) 動機: 一般ドキュメント(TEXT, HTML, EXCEL, EAMIL, WORD, PDF)には, 多くの位置情報, たとえば住所や地名が記載されている. これらの位置情報は人間にとって可読性・伝達性が高い. 従来の GIS では, 緯度経度のような座標値といった, 機械に扱いやすい位置情報を主対象としてきた. その結果, 多くの位置情報を含む一般ドキュメントは, GIS の対象の中には入っていなかった.

(2) アプローチ: 座標のように機械に向く位置情報を直接位置参照情報, 一方, 住所や地名のように, 人間にやさしい位置情報を間接位置参照情報と呼ぶ. 間接位置参照情報を直接位置参照情報へ変換する処理をジオコーディングと呼ぶ. 本研究では, ジオコーディングにより, 一般ドキュメントをパソコン上で簡単に地図と連携させて扱える枠組みとして「空間ドキュメント管理システム」を提案・実現した. 従来, 専門家のツールであった GIS と違い, これを用いることにより日常的に利用しているドキュメントを空間的に管理できる.

(3) 意義: サーチエンジン出現の前は, 情報検索の対象は十分に構造化されたデータだけであったが, 現在では情報検索の対象は一般ドキュメントである. 同様に, 現在の GIS は構造化された空間データだけを対象としているが, 提案システムでは, 十分に構造化してないドキュメントを対象とする.

(4) 特徴:
- 一般ドキュメントを POI (Point of Interest) に変換する. ドキュメントから住所抽出を行い, アドレスマッチングにより, 緯度経度に変換して POI を生成する. ドラック&ドロップ操作で上記の処理を実行可能.
- ドキュメントを単位にしたレイヤ管理.
- テキスト検索と空間検索の組み合せ利用.
- 全国のアドレスマッチング(街区レベル)と全国の 25,000 分の 1 の背景図を提供する.
- POI を shape と G-XML 2.0 の形式で出力可能.

(5) その他:
- 無料で一般公開する予定. 現在, データ利用の許可のために準備中.
- 本研究は, 厚生労働科学研究費補助金(健康科学総合研究事業)「地理及び社会状況を加味した地域分析方法の開発に関する研究」(代表:浅見泰司)の支援を受けている.

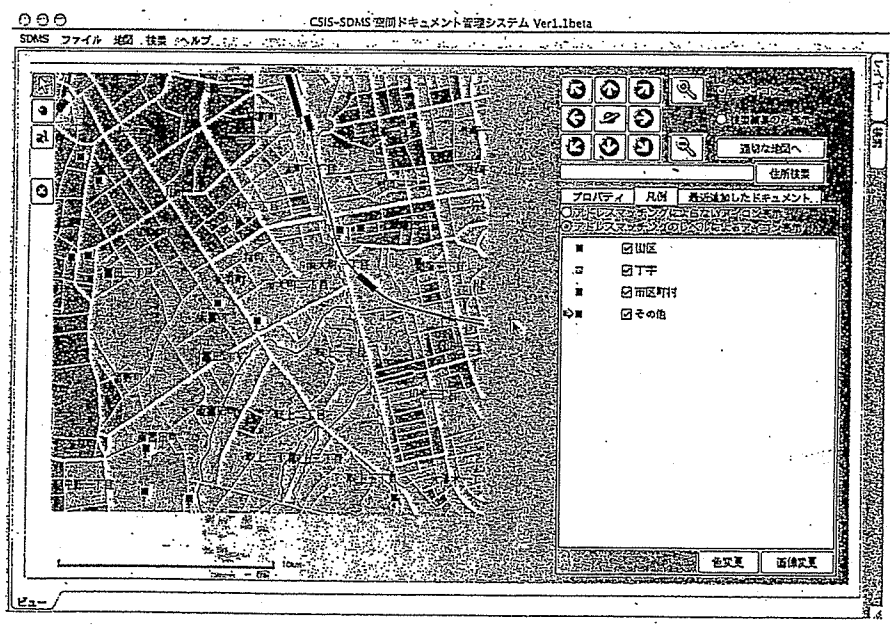


図 1: 空間ドキュメント管理システムの画面の例
あるテキストファイルの中に住所のテーブルが含まれていた場合, そのテキストファイルを空間ドキュメント管理システムにドラック&ドロップするだけで, 含まれる住所記述を地図上の点として抽出し, 表示できる. 図では, 住所がマッチングできたレベルにより, 点の色を違えて表示している.

動物園の土壌細菌叢と薬剤耐性菌分布の調査

産業医科大学・医学部・微生物学

福田和正、池野貴子、郡山一明、谷口初美

薬剤耐性菌の分布と抗菌物質産生菌の探索　宮脇、池野、福田他

日本細菌学会九州支部総会　2005年7月8日

ごみとDNA ―土壌細菌叢の解析―　　谷口初美　化学療法の領域

22, 4 (2006)　化学療法学会誌

動物園の土壌細菌叢と薬剤耐性菌分布の調査

産業医科大学・医学部・微生物学
福田和正、池野貴子、郡山一明、谷口初美

薬剤耐性菌の分布と抗菌物質産生菌の探索　宮脇、池野、福田他　日本細菌学会九州支部総会
2005年7月8日
ごみとDNA ―土壌細菌叢の解析―　　谷口初美　化学療法の領域　　22, 4 (2006) 化学療法学会誌

本研究は、自然環境におけるバイオハザード評価のために、微生物学的基礎データを収集することを目的とする。自然環境における微生物の主たる供給源は土壌と考えられる。そこで自然環境におけるバイオハザードを評価するためには、その供給源である土壌の細菌叢を理解することが重要であると考える。当該研究室では廃棄物処分場の硫化水素ガス発生対策のために、平成14年度より廃棄物処分場の土壌の細菌叢の網羅的検査法を構築してきた。本研究ではこの手法を応用して、土壌細菌叢のデータを収集する。また感染症において薬剤耐性菌の問題は深刻である。そこで、土壌細菌叢の中で薬剤耐性菌の占める割合についても調査を行う。

平成16年度は廃棄物処分場、山、畑、大学構内など理化学性状の異なる土壌について、好気培養で増殖する菌のうちの薬剤耐性菌の占める割合を調査した。その結果、廃棄物処分場、山、畑の土壌に比べ、病院を併設する大学構内の土壌において、薬剤耐性菌の占める割合が大きいことが明らかとなった。とくにアミノベンジルペニシリンについての耐性菌の検出率が高かった。この理由として、病院におけるペニシリン系抗菌薬の多用、またはヒトが常時住む場所であることが可能性として考えられた。

そこで平成17年度は、病院施設に次いで抗菌薬を多用する可能性が高い動物展示施設の土壌について、前年度同様に好気培養で増殖する菌のうち薬剤耐性菌が占める割合を調べた。また母数として、全菌数を蛍光染色法で計測し、好気培養で増殖する好気性菌、通性嫌気性菌の菌数も測定した。

Ⅰ．検査材料と方法

1：土壌サンプル

2005年9月、動物展示施設の15箇所19サンプルの表層土壌を採取した。ふれあい広場3箇所（ZS3、ZS4、ZS5）、広場2箇所（ZC1、ZC2）、動物展示小屋5箇所（ZS6、ZS7、ZS9、ZS10、

ZS11)、鳥展示小屋４箇所 ZS1、ZS2、ZS8、ZS12)、ウサギ展示小屋３箇所（ZR1、ZR2、ZR3）と消毒後１箇所（ZS13)、ウサギ小屋近くの土壌１箇所（ZRCont)。

２：全菌数

EtBr 蛍光染色法で計測した。土壌希釈液をフィルターをセットしたろ過器にバファーを加え、さらに EtBr 水溶液を加える（終濃度 100μg／l)。室温で 10 分間放置後、吸引ろ過。ろ過滅菌水 3ml でフィルターを洗浄。無蛍光オイルで封入。検鏡計測。

３：好気培養による菌数計測

土壌懸濁液の 10 倍希釈系列100μ 1 を環境微生物用の標準寒天培地２枚づつに接種した後、30℃において培養を行い、6 日後に菌数を測定した。

４：薬剤耐性菌

４－１）薬剤含有培地の作製

薬剤として ABPC (アミノベンジルペニシリン)、 SM (ストレプトマイシン)、EM (エリスロマイシン)、TC (テトラサイクリン)、OFLX (オフロキサシン) を用いた。環境細菌用標準培地を滅菌 (オートクレーブ) した後、終濃度がそれぞれ 100 μg/ml、10 μg /ml、1 μg /ml となるよう添加した。

４－２）土壌微生物の培養方法

土壌希釈液 0. 1ml を各濃度の薬剤入り培地に接種後、30 ℃で 40 時間培養した。対象群として薬剤の入っていない培地にも土壌希釈液 0. 1ml を接種し、培養 40 時間後に各薬剤培地及び薬剤不含有培地のコロニー数を計測した。

４－３）薬剤耐性菌の検出率

40 時間培養後の薬剤耐性菌のコロニー数を、薬剤不含有培地のコロニー数を母数として割合を算出した。

Ⅱ．結果：

１：全菌数と好気培養による菌数計測

全菌数はいずれの土壌でも 1g 当たり 10^9～10^{10}個であった。この結果は、処分場や山土、大学構内、畑などの結果と同じであった。

好気培養法で増殖する菌は、15 サンプルでは全菌数の 1～10%程度であった。しかし ZS1, ZS3, ZS5, ZS12 の 4 サンプルでは、それぞれ約 50%, 30%, 30%, 80%で、特に ZS12 は高い値を示した。

２：薬剤耐性菌

２－１：土壌別の薬剤耐性菌の検出率

・広場、ふれあい、動物飼育小屋に比べ、鳥飼育小屋で耐性菌の検出率が高い傾向が見られた。ノウサギ小屋で消毒後 (ZS13) に検出率の顕著な増加が見られた。（図１）

- 各薬剤 1μg/ml については、図 2 にまとめた。ABPC, OFLX, EM. SM, TC の順に耐性菌が多かった。薬剤別に耐性菌（1μg/ml 耐性）の割合が 20%以上を超える土壌の数は、ABPC 耐性を示す土壌が 12 箇所, OFLX 9 箇所, EM 7 箇所, SM 3 箇所, TC 0 箇所であった。
- 各薬剤 10μg/ml については、図 3 にまとめた。
- 各薬剤 100μg/ml に対する耐性菌はほとんど検出されなかった。

図1:土壌別の耐性菌検出率

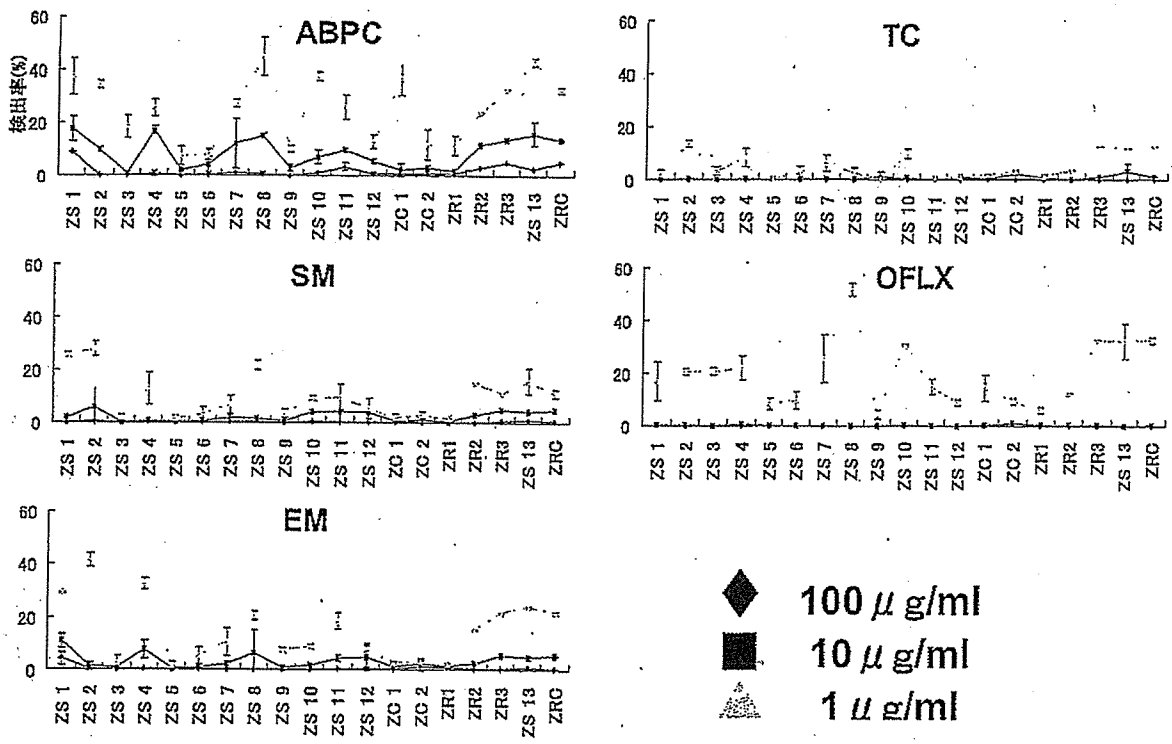
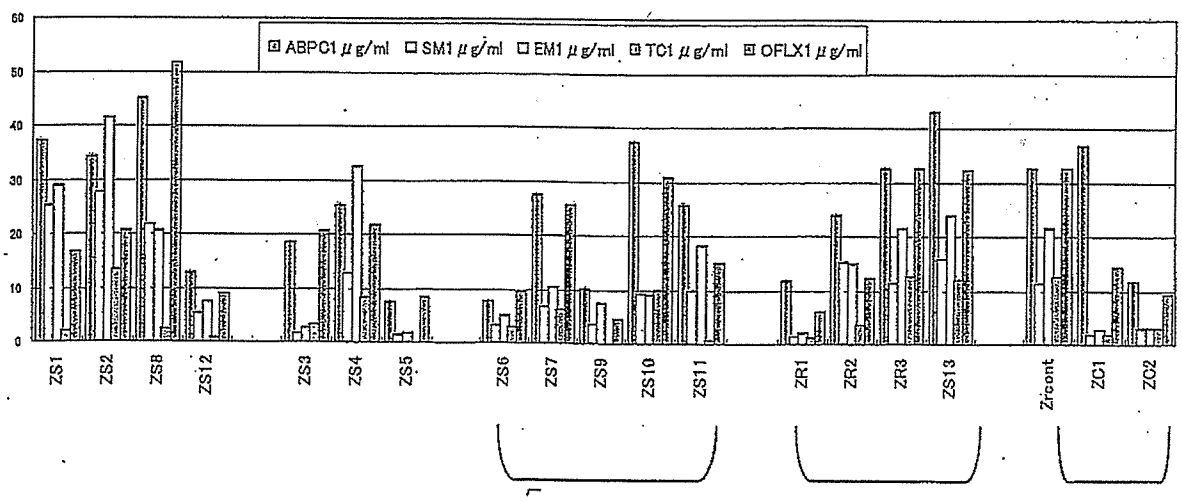


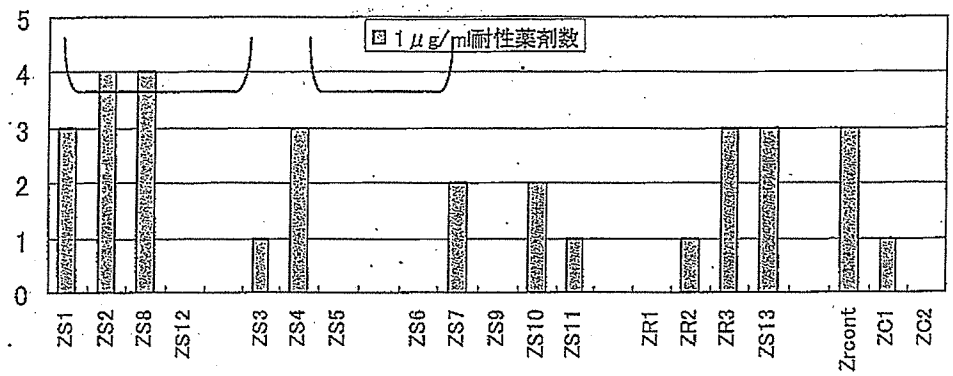
図 2：各薬剤 1μg/ml に対する耐性菌検出率



鳥　　ふれあい　　動物　　ノウサギ　　広場

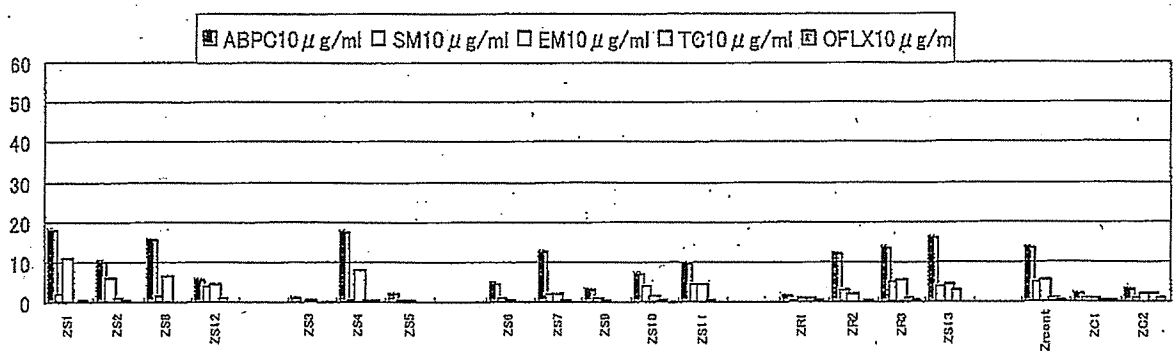
2－2：土壌別の耐性菌の割合が20%を超える薬剤の数（各薬剤1μg/ml）

鳥展示小屋4箇所	0, 3, 4, 4	(ZS12, ZS1, ZS2, ZS8)
ふれあい広場3箇所	0, 1, 3	(ZS5, ZS3, ZS4)
動物展示小屋5箇所	0, 0, 1, 2, 2	(ZS6, ZS9, ZS11, ZS7, ZS10)
ノウサギ小屋4箇所	0, 1, 3, 3	(ZR1, ZR2, ZR3, ZS13)
広場など3箇所	0, 1, 3	(ZC2, ZC1, ZRcont)



・1剤耐性はABPCに、2剤耐性はABPCとOFLXに、3剤耐性はABPC,OFLX,EMに耐性を示した。

図3：各薬剤10μg/mlに対する耐性菌検出率



Ⅲ．まとめ

- 全菌数はいずれの土壌でも他の表層土壌と同様に10^9～10^{10}/gであった。これは通常土壌の1g当たりの全菌数は、ほぼ一定に保たれているという、これまでの結果と一致するものであった。
- 好気培養法で増殖する菌数は、15サンプルで全菌数の1～10%程度であり、これも

- 通常土壌で得られる結果と一致していた。しかし、4サンプルでその割合が高い傾向が見られ、特に鳥類飼育小屋の ZS12 はその割合が全菌数の約 80%と他の土壌に比べて異常に高かった。原因として湿った土壌であったことが可能性として考えられた。感染症との関連は今後の検討課題である。
- 耐性菌の検出率は、土壌ごとに違いが見られた。
- 特に鳥類飼育小屋の ZS1, 2, 8 は TC 以外の 4 剤の耐性率が他のサンプルに比べ高い傾向であった。しかし鳥類飼育小屋であっても、ZS12 の場合は耐性率が低く、異なる傾向を示した。ZS12 は好気培養で増殖する菌の割合が異常に高く、母数が大きいために検出率が低かった可能性が考えられた。
- ノウサギ展示小屋の消毒前後(ZR1, 2, 13)を比較すると、消毒後の ZR13 は ABPC に耐性の割合が増加していた。消毒による菌種の選択が起きた可能性が考えられる。回復までの経時的検査が必要である。
- 耐性の割合は、昨年同様、ペニシリン系薬剤であるアミノベンジルペニシリンに対する耐性菌の検出率が最も高かった。これは 1μg/ml 耐性に顕著に見られた。グラム染色の結果、土壌の細菌はグラム陽性菌が多いにもかかわらず、ABPC 耐性菌の場合はグラム陰性桿菌が多かった。臨床上 ABPC 耐性は 8μg/ml 以上とされている。10μg/ml 以上の耐性を示す菌が 10%以上生息する土壌が 10 箇所であった(図 3)。自然界には ABPC 自然耐性の菌が多く生息している可能性と、本邦ではペニシリン系薬剤の使用頻度が最も高いことを反映している可能性も考えられた。
- ABPC の次に、OFLX, EM, SM の順に耐性率が高く、TC は最も低かった。

Ⅳ. 考察

動物展示場は、昨年の病院を併設している大学構内の土壌と同様に、耐性菌の占める割合が高かった。今回は 9 月の暑い時期に採取した、1 回だけの結果であるので、次年度は寒い時期のサンプル採取を行い、季節による違いを見る。また薬剤耐性菌だけでなく、菌種、菌叢の解析も行う予定である。

International Journal of Health Geographics

Methodology

A flexibly shaped spatial scan statistic for detecting clusters

Toshiro Tango and Kunihiko Takahashi

18 MAY 2005

BioMed Central

UK

International Journal of Health Geographics

BioMed Central

Methodology

Open Access

A flexibly shaped spatial scan statistic for detecting clusters

Toshiro Tango* and Kunihiko Takahashi

Address: Department of Technology Assessment and Biostatistics, National Institute of Public Health, 3–6 Minami 2 chome Wako, Saitama 351-0197 Japan

Email: Toshiro Tango* - tango@niph.go.jp; Kunihiko Takahashi - kunihiko@niph.go.jp

* Corresponding author

Published: 18 May 2005

Received: 14 April 2005

Accepted: 18 May 2005

International Journal of Health Geographics 2005, 4:11 doi:10.1186/1476-072X-4-11

This article is available from: http://www.ij-healthgeographics.com/content/4/1/11

Abstract

Background: The spatial scan statistic proposed by Kulldorff has been applied to a wide variety of epidemiological studies for cluster detection. This scan statistic, however, uses a circular window to define the potential cluster areas and thus has difficulty in correctly detecting actual noncircular clusters. A recent proposal by Duczmal and Assunção for detecting noncircular clusters is shown to detect a cluster of very irregular shape that is much larger than the true cluster in our experiences.

Methods: We propose a flexibly shaped spatial scan statistic that can detect irregular shaped clusters within relatively small neighborhoods of each region. The performance of the proposed spatial scan statistic is compared to that of Kulldorff's circular spatial scan statistic with Monte Carlo simulation by considering several circular and noncircular hot-spot cluster models. For comparison, we also propose a new bivariate power distribution classified by the number of regions detected as the most likely cluster and the number of hot-spot regions included in the most likely cluster.

Results: The circular spatial scan statistics shows a high level of accuracy in detecting circular clusters exactly. The proposed spatial scan statistic is shown to have good usual powers plus the ability to detect the noncircular hot-spot clusters more accurately than the circular one.

Conclusion: The proposed spatial scan statistic is shown to work well for small to moderate cluster size, up to say 30. For larger cluster sizes, the method is not practically feasible and a more efficient algorithm is needed.

Background

The question of whether disease cases are clustered in space has received considerable attention in the literature [1-4]. Although many statistical tests for disease clusters have been proposed, most tests suffer from multiple testing problems due to one or two unknown parameters that must be set prior to their applications. For example, Cuzick and Edwards's procedure [5] has an unknown number k of nearest-neighbours and Besag and Newell's method [6] has an unknown number of cases k for the size of the cluster. As far as we know, the spatial scan statistic proposed by Kulldorff [7,8] and Tango's maximized excess events test [9,10] are exceptions and take multiple testing into account in the sense that we have only to specify the maximum possible cluster size. Especially, Kulldorff's circular spatial scan statistic has been applied to a wide

variety of epidemiological studies for cluster detection (for example, see [11-13]). In recent power comparisons of disease clustering tests, his scan statistic has been shown to be the most powerful for detecting localized clusters [14,15]. It should be noted, however, 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 detecting the true cluster correctly is a different matter.

As the circular spatial scan statistic uses a "circular window" with variable size to define the potential cluster area, it is difficult to correctly detect noncircular clusters such as those along a river. Most geographical areas are noncircular. Furthermore, in our experience in applying SaTScan program [16] to various data, even if the null hypothesis is rejected, the circular spatial scan statistic tends to detect a larger cluster than the true cluster by absorbing surrounding regions where there is no elevated risk. It should be noted that although Kulldorff originally made no assumptions about the shape of the scanning window in his paper [8], a circular scanning window has been used in almost all purely spatial applications especially for the availability of software and computational speed.

Recently, Patil and Taillie [17] and Duczmal and Assunção [18] proposed non-circular spatial scan statistics based on the likelihood ratio test formulated in the same way as in the circular spatial scan statistic. To avoid undertaking computationally infeasible searches, they considered different approaches. Patil and Taillie [17] used the notion of "upper level set" to reduce the size of windows to be scanned and proposed "upper level set scan statistic". However, they do not discuss how to select the level g which defines the upper level set and do not provide any illustrations of their method nor any results of comparison with the circular scan statistic. Duczmal and Assunção [18], on the other hand, have applied a simulated annealing method in which they try to examine only the most promising windows using a graph-based algorithm to obtain the local maxima of a certain likelihood function over a subset of the collection of all the connected regions. Their method seems to be very complicated but they do not show any programmable procedure of their method. In our experience using their program (personal communication to Professor Duczmal via email) which is executable with the Borland C++ Builder 6, their scan statistic, in most cases, detected a cluster of peculiar shape that was much larger than the true cluster by absorbing not only surrounding regions with non-elevated risk but also faraway regions with non-elevated risk. An example of such properties of Duczmal and Assunção's procedure is shown later in comparison with the circular spatial scan statistic and the proposed flexible spatial scan statistic. That is why we did not include both the Patil and Taille method and Duczmal and Assunção's procedure in our simulation for comparison.

In this paper, we propose an alternative *flexibly shaped spatial scan statistic* ('flexible spatial scan statistic' hereafter) in which the detected cluster is allowed to be flexible in shape while at the same time the cluster is confined within relatively small neighborhoods of each region. The performance of the flexible spatial scan statistic is compared with that of the circular spatial scan statistic using Monte Carlo simulation. In comparing performance we examined not only the usual power but also the newly introduced bivariate power distribution classified by the number of regions detected as the most likely cluster and the number of hot-spot regions included in the most likely cluster. The proposed flexible spatial scan statistic is illustrated with some simulated disease maps for the Tokyo Metropolitan area.

Methods

Consider the situation where an entire study area is divided into m regions (for example, county, enumeration districts, etcetera). The number of cases in the region i is denoted by the random variable N_i with observed value n_i, $i = 1, ..., m$. Under the null hypothesis H_0 of no clustering, the N_i are independent Poisson variables such that

$$H_0 : E(N_i) = \xi_i,\ N_i \sim \text{Pois}(\xi_i),\ i = 1, ..., m \quad (1)$$

where Pois(e) denotes Poisson distribution with mean e and the ξ_i are the null expected number of cases in the region i. To specify the geographical position of each region, we will use the coordinates of the administrative population centroid.

Under this situation, the circular spatial scan statistic imposes a circular window Z on each centroid. For any of those centroids, the radius of the circle varies from zero to a pre-set maximum distance d or a pre-set maximum number of regions K to be included in the cluster. If the window contains the centroid of a region, then that whole region is included in the window. In total, a very large number of different but overlapping circular windows are created, each with a different location and size, and each being a potential cluster. Let Z_{ik}, $k = 1,..., K$, denote the window composed by the $(k - 1)$-nearest neighbours to region i. Then, all the windows to be scanned by the circular spatial scan statistic are included in the set

$$Z_1 = \{Z_{ik} \mid 1 \le i \le m, 1 \le k \le K\} \quad (2)$$

A flexible scan statistic we propose, on the other hand, imposes an *irregularly shaped* window Z on each region by connecting its adjacent regions. For any given region i, we

create the set of irregularly shaped windows with *length k* consisting of k connected regions including i and let k moves from 1 to the pre-set maximum K. To avoid detecting a cluster of *unlikely peculiar shape*, the connected regions are restricted as the subsets of the set of regions i and $(K - 1)$-nearest neighbours to the region i where K is a pre-specified maximum length of cluster. In total, as in the circular spatial scan statistic, a very large number of different but overlapping arbitrarily shaped windows are created. Let $Z_{ik(j)}, j = 1,..., j_{ik}$ denote the j-th window which is a set of k regions connected starting from the region i, where j_{ik} is the number of j satisfying $Z_{ik(j)} \subseteq Z_{ik}$ for $k = 1,..., K$. Then, all the windows to be scanned are included in the set

$$Z_2 = \{Z_{ik(j)} \mid 1 \le i \le m, 1 \le k \le K, 1 \le j \le j_{ik}\} \quad (3)$$

In other words, for any given region i, the circular spatial scan statistic consider K concentric circles, whereas the flexible scan statistic consider K concentric circles plus all the sets of connected regions (including the single region i) whose centroids are located within the K-th largest concentric circle. So, the size of Z_2 is far larger than that of Z_1 which is at most mK. Details of the algorithm that we adopted to find all these arbitrarily shaped windows within a pre-specified maximum length K are given in the Appendix.

Under the alternative hypothesis, there is at least one window Z for which the underlying risk is higher inside the window when compared with outside. In other words, we are considering the following hypothesis:

$$H_0 : E(N(Z)) = \xi(Z), \text{ for all } Z, \; H_1 : E(N(Z)) > \xi(Z), \text{ for some } Z \quad (4)$$

where $N()$ and $\xi()$ denote the random number of cases and the null expected number of cases within the specified window, respectively. For each window, it is possible to compute the likelihood to observe the observed number of cases within and outside the window, respectively. Under the Poisson assumption, the test statistic, which was constructed with the likelihood ratio test [8], is given by

$$\sup_{Z \in Z} \left(\frac{n(Z)}{\xi(Z)} \right)^{n(Z)} \left(\frac{n(Z^c)}{\xi(Z^c)} \right)^{n(Z^c)} I\left(\frac{n(Z)}{\xi(Z)} > \frac{n(Z^c)}{\xi(Z^c)} \right) \quad (5)$$

where Z^c indicates all the regions outside the window Z, and $n()$ denotes the observed number of cases within the specified window and $I()$ is the indicator function. The window Z^* that attains the maximum likelihood is defined as the *most likely cluster* (MLC). To find the distribution of the test statistic under the null hypothesis, Monte Carlo hypothesis testing [19] is required. In this paper, p-value of the test is based upon the null distribution of likelihood ratio test statistic with a large number (we used 999) of Monte Carlo replications of the data set generated under the null hypothesis. It should be noted that, in the same manner as the circular spatial scan statistic, the flexible spatial scan statistic is also able to locate secondary clusters that do not overlap the most likely cluster but are still statistically significant.

Results

Illustrations and powers

In this section, we will compare the flexible spatial scan statistic with the circular spatial scan statistic. As an entire study population, we will use $m = 113$ regions comprising the wards, cities and villages in the area of Tokyo Metropolis and Kanagawa prefecture in Japan (Figure 1). The variability of regional populations for $m = 113$ regions is: 25 percentile = 56, 704, median = 142, 320 and 75 percentile = 200, 936.

Hot-spot clusters

We will consider the following four hot-spot clusters where the expected total number of cases $\sum_{i=1}^{m} \xi_i$ is set to be 200 under the null hypothesis.

1. Cluster A = {14, 15, 20}

2. Cluster B = {14, 15, 20, 26}

3. Cluster C = {14, 15, 26, 27}

4. Cluster D = {73, 74, 75, 76, 78}

where the region included in a hot-spot cluster is called a "hot-spot region" (hot-spot region numbers are shown in Figure 1). The relative risk within any cluster R is set to three, i.e.,

$$H_1 : N(R) \sim \text{Pois}(\theta\xi(R)), \; \theta = 3.0 \quad (6)$$

The cluster A is considered here as an example of a circular cluster that can be in the set of the circular windows and is expected to be identified by the circular spatial scan statistic more often than by the flexible spatial scan statistic. The other clusters are examples of noncircular clusters that are not in the set of the circular windows and thus cannot be identified correctly by the circular spatial scan statistics. For example, consider the region $i_0 = 15$ as the starting region and the set of $(K - 1)$-nearest neighbours to the region 15, which is listed as follows in the ascending order of distance from the region 15:

15, 14, 20, 12, 4, 26, 13, 27, 16, 40, 19, 42, 10,...,

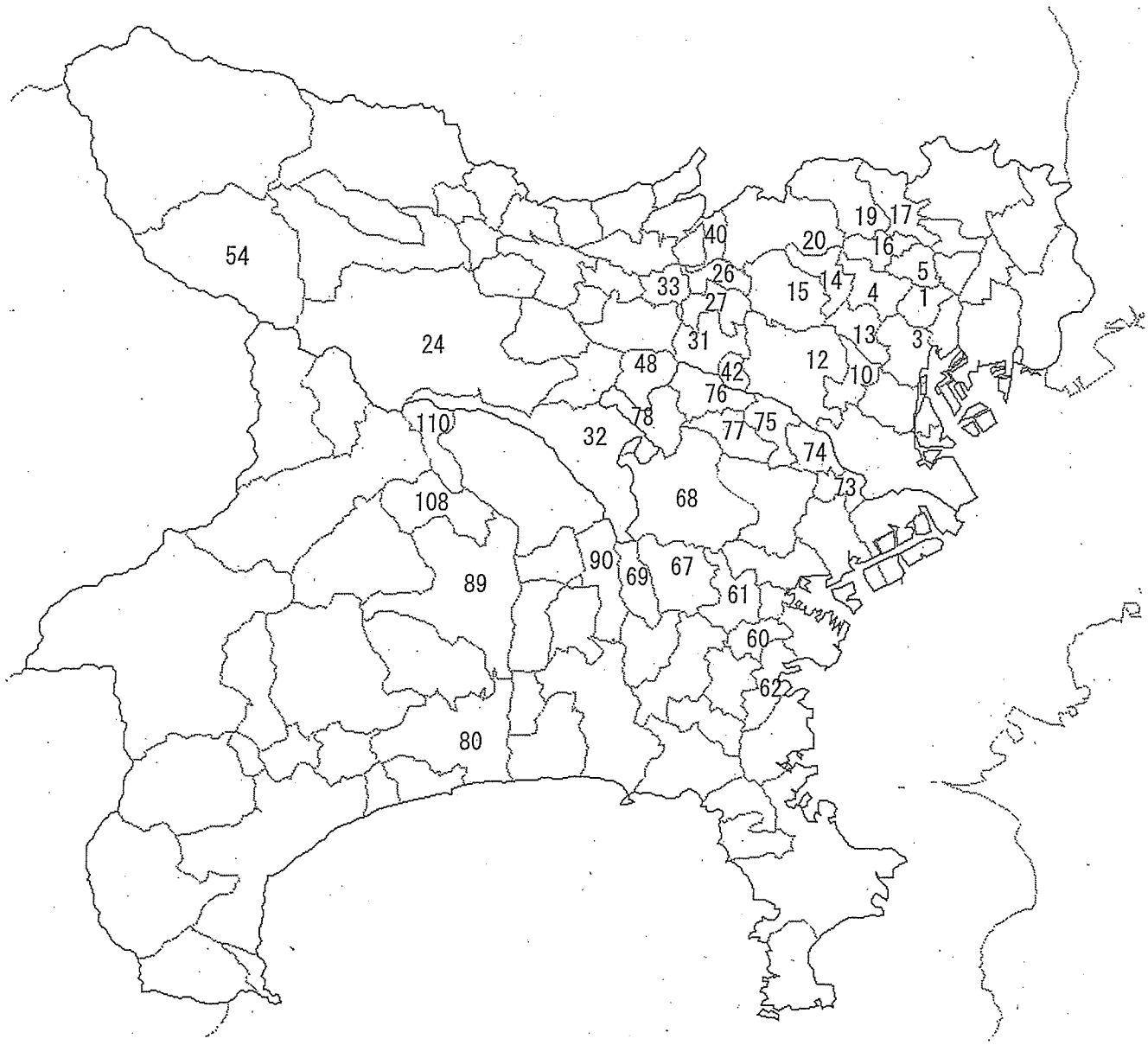


Figure 1
An entire study population for simulation studies. The 113 regions comprising wards, cities and villages in the area of Tokyo Metropolis and Kanagawa prefecture in Japan. The region number used in the text is shown. Especially, The region numbers of four hot-spot clusters **A-D** are **A** = {14, 15, 20}, **B** = {14, 15, 20, 26}, **C** = {14, 15, 26, 27}, and **D** = {73, 74, 75, 76, 78}, respectively.

In this case, circular windows are {15}, {15, 14}, {15, 14, 20}, {15, 14, 20, 12}, ... When the starting region is 14 or 20, the corresponding set of $(K - 1)$-nearest neighbours is

14, 15, 20, 4, 16, 13, 19, 12, 5, 1, 17, 10, 26, 3, 27,...,

and

20, 14, 15, 19, 16, 4, 17, 26, 40, 13, 5, 12, 1, 27,...,

respectively. In both cases, cluster B and C are easily found to be not in the set of circular windows. The cluster **D** is considered as an example of a long and narrow cluster as is shown in Figure 1.

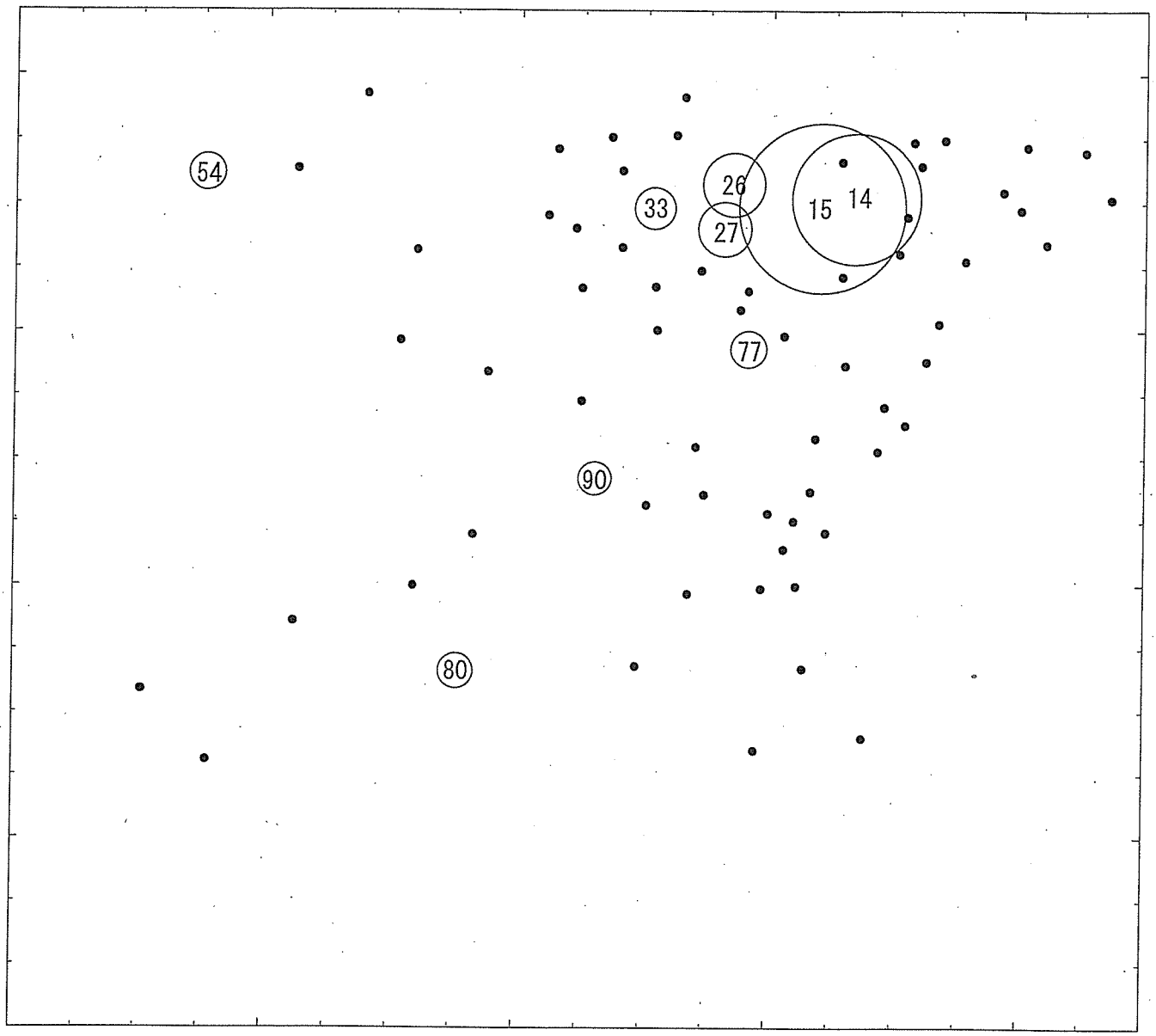


Figure 2
A random sample from cluster model C. Dots describe the centroids of regions with some cases. Circles are drawn only for the regions whose standardized risk ratios are statistically significantly larger than 1 at $\alpha = 0.05$ and the region number is placed in stead of dot. The radius is set inversely proportional to the tail probability.

Illustrative example

As an illustration, we will apply the circular spatial scan statistic, the flexible spatial scan statistic and Duczmal and Assunção's spatial scan statistic to the disease map shown in Figure 2 which is a random sample of $n = 235$ cases assuming the cluster model C. Circles are drawn only for the regions whose observed-expected ratio (standardized risk ratio) is statistically significantly larger than 1 at $\alpha = 0.05$. The radius of the circles is set inversely proportional to the upper tail p-value. The number shown in Figure 2 indicates the region number. Figure 2 obviously suggests the clusters occurring in the area including regions $\{14, 15, 26, 27, 33\}$.

Before applying the three spatial scan statistics, we have to specify a common maximum length K for the most likely cluster. This makes comparisons to a certain extent fair. In this example, we chose two kinds of maximum length $K =$

15 and $K = 20$ since it is not unreasonable to assume that an actual cluster size will be less than one third or one fourth of the size of the whole study area.

Irrespective of the value of K, the circular spatial scan statistic detected the regions {14, 15} as MLC with log likelihood ratio = 20.1, $p = 1/(999 + 1) = 0.001$ and the estimated relative risk is $\hat{\theta} = 3.47$. This is shown in Figure 3(a). The flexible spatial scan statistic, regardless of the value K, detected the regions {14, 15, 26, 27, 33} as MLC with log likelihood ratio = 29.7, $p = 0.001$ and the estimated relative risk is $\hat{\theta} = 3.41$. This is shown in Figure 3(b). Duczmal and Assunção's method, on the other hand, detected a cluster of peculiar shape that is much larger than the true cluster. In the case of $K = 15$, their scan statistic detected an area consisting of $K = 15$ connected regions {14, 15, 24, 26, 27, 31, 32, 33, 48, 54, 69, 77, 78, 90, 110 } as MLC with log likelihood ratio = 31.8, $p = 0.001$ and the estimated relative risk is $\hat{\theta} = 2.40$. This is shown in Figure 4(a). Figure 4(b) shows the most likely cluster {14, 15, 26, 27, 31, 32, 33, 48, 60, 61, 62, 67, 69, 77, 78, 80, 89, 90, 108, 110 } detected by Duczmal and Assunção's scan statistic for $K = 20$ where the length of MLC is also the same as $K = 20$ and log likelihood ratio = 36.0, $p = 0.001$ and the estimated relative risk is $\hat{\theta} = 2.26$. In the case of $K = 15$, the results of the three scan statistics are summarized in Table 1. Although the most likely cluster detected by Duczmal and Assunção's scan statistic has the largest log likelihood ratio among three scan statistics, it has detected MLC surprisingly larger than the true cluster.

Using a PC(Windows XP, CPU pentium 4, 3.2 GHz), the execution time of the flexible spatial scan statistic in this example is 14 seconds for $K = 15$ and 379 seconds for $K = 20$ which is certainly greater than that for the circular spatial scan statistic (less than 1 second for both $K = 15$ and $K = 20$).

Power comparison

In the power comparison, we chose $K = 15$. To compare the power of the flexible spatial scan statistic with that of the circular spatial scan statistic based upon Monte Carlo simulation, we will introduce a new bivariate power distribution $P(l, s)$ classified by the *length* l of the significant MLC and the number s of hot-spot regions included in the most likely cluster:

$$P(l,s) = \frac{\#\{\text{significant MLC has length } l \text{ and includes } s \text{ hot-spot regions}\}}{\#\{\text{trials for each simulation}\}}, \quad (7)$$

where $l \geq 1$ and $s \geq 0$. Based on $P(l, s)$, we examined the following powers,

1. the usual power, i.e., $P(+,+) = \sum_{l\geq 1} \sum_{s\geq 0} P(l, s)$,

2. the joint power $P(l, s)$, especially $P(s^*, s^*)$ where s^* is the length of the hot-spot cluster assumed in the simulation.

3. the marginal power distribution of $s(\geq 0)$, $P(+, s) = \sum_{l\geq 1} P(l, s)$ and its conditional power $P(+, s)/P(+,+)$,

4. the marginal power distribution of $l(\geq 1)$, $P(l, +) = \sum_{s\geq 0} P(l, s)$.

The powers are calculated for tests of nominal α levels of 0.05 and for the expected total number of cases 200 under the null hypothesis, which are based on Monte Carlo simulation using Poisson random numbers. For each simulation, 1,000 trials were carried out. The resultant power distribution $P(l, s) \times 1000$ is shown in Tables 2, 3, 4, 5 for each of the four cluster models, respectively, in the form of cross table classified by l ("*length*" in tables) and s ("*include*" in tables).

1) Usual power

Both tests have the same size 0.043 (distribution of length of significant MLC is omitted) and are shown to have high powers for the hot-spot clusters considered here. The flexible spatial scan statistic generally has higher power except for the model A (circular cluster) where, however, the difference is small.

2) Joint powers at (s, s*) and at its neighbours*

Table 2 shows the good characteristics of the circular spatial scan statistic. Namely, the circle-based scan statistic could detect circular hot-spot cluster A with length $s^* = 3$ considerably more accurately with power 738/1000 compared to 142/1000 of the flexible spatial scan statistic. Tables 3, 4, 5, on the other hand, show that the power of the circular spatial scan statistic in detecting exactly non-circular hot-spot clusters is 0/1000 due to the circular window. However, the circular spatial scan statistic is seen to be able to include some of the hot-spot regions into MLC reasonably well. For example, when applied to the noncircular cluster B with length $s^* = 4$, three or four regions including three hot-spot regions can be detected as the most likely cluster with relatively high power (523 + 65)/1000 = 0.588 (Table 3). When applied to the model D with length $s^* = 5$, the similar high power 363/1000 can be observed at $(l, s) = (6, 4)$ (Table 5). The flexible spatial scan statistic, on the other hand, has no such high power at a single point (l, s) near (s^*, s^*). However, the characteristic of the flexible spatial scan statistic is that the support of the power distribution is distributed in a relatively narrow range of l on the line $s = s^*$, i.e, we have $s^* \leq l \leq 12$ in the four cluster models considered here.

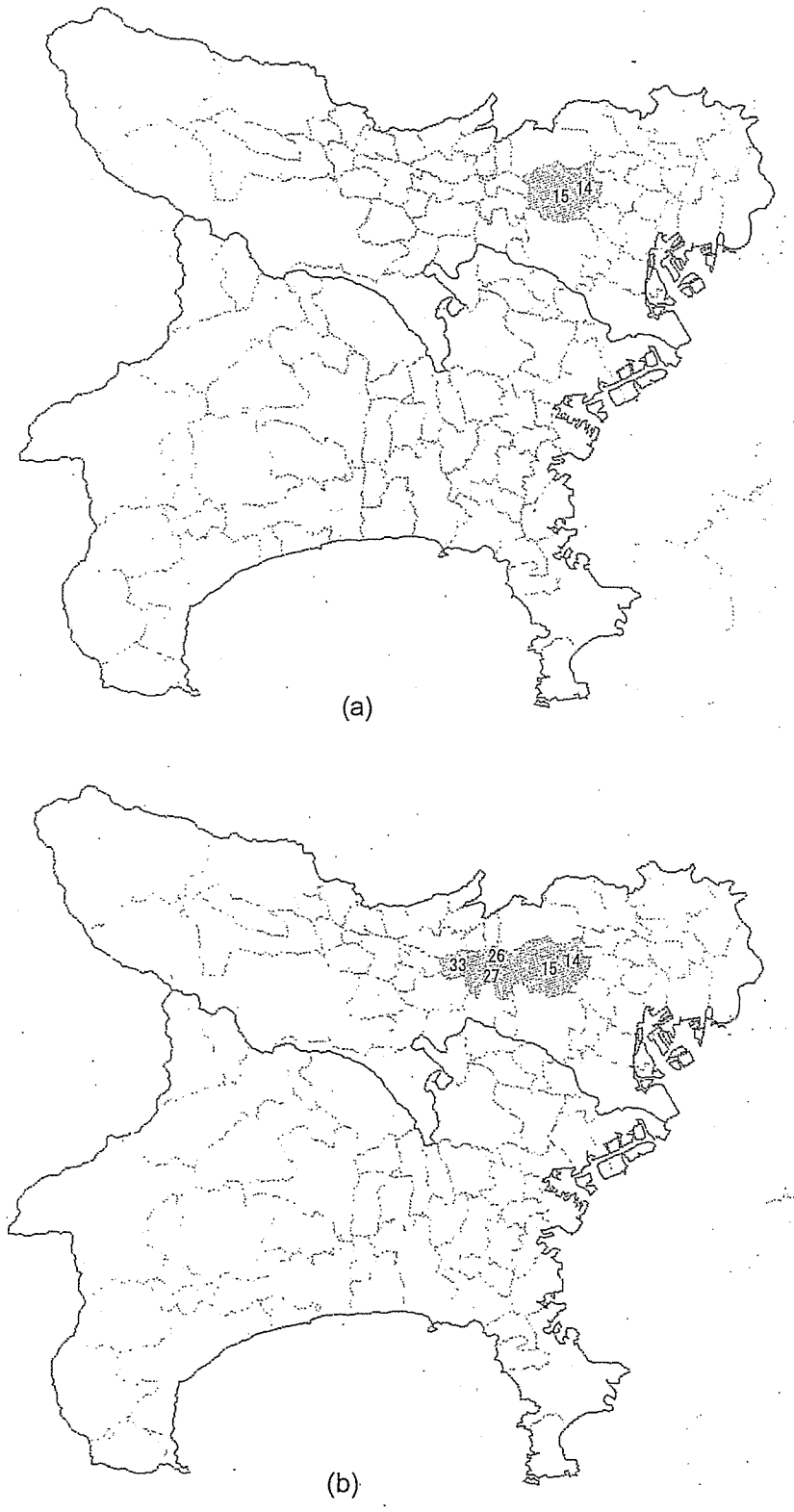


Figure 3
The most likely cluster detected by the circular and the flexible spatial scan statistic. (a) Detected by the circular spatial scan statistic for both $K = 15$ and $K = 20$ and (b) by the flexible spatial scan statistic for both $K = 15$ and $K = 20$, when applied to a random sample from the cluster model $\mathbf{C} = \{14, 15, 26, 27\}$.