Methods for Automating the Exploration of the Geocyberspace of Health and Crime Data

Reviving the Geographical Analysis Machine

Stan Openshaw and Max Craglia

Universities of Leeds and Sheffield

Contents

- Background
- Historical Aspects
- The IARC Review of Spatial Clustering Methods
- How GAM works
- GAM on crime data
- Other GAMs
- Conclusions

Background

- **There is a VAST explosion of Geographically Referenced Data due to**
  - developments in IT
  - GIS revolution
  - computerization of everything
  - falling costs of storage
  - dramatic changes in the price-performance of computing

There is a VAST and RAPIDLY

growing

GEOCYBERSPACE of information

- this geocyberspace constitutes the **RAW MATERIALS** from which new knowledge, new concepts, and new discoveries etc will have to be extracted in the 21st Century
- **most is geographical information** although the geography is often ignored in data mining

Many Spatial Data Bases are now available for analysis

- BUT

- very few suitable spatial analysis tools exist that can cope with <u>BOTH</u>

  the data

  and

  the users

The NEED is for Exploratory Spatial Data Analysis

- user's have few if any hypotheses to test
- most end-users are NOT academics but practitioners interested in pursuing APPLIED goals and not research
- there are MANY data sets that NEED to be analyzed
- an INCREASING NUMBER have a strong imperative for analysis once they exist

Many users now have a requirement for

DATA EFFICIENT and USER-FRIENDLY

exploratory spatial data analysis methods capable of being safely used by people who do not have higher degrees in the statistical or spatial sciences

User Friendly Spatial Analysis:

- provides analyses that users need
- simple to perform
- highly automated making it fast and efficient
- readily understood
- results are self-evident and can be communicated to non-experts
- safe and trustworthy

These Methods can be classified as follows

a blank slide... few if any exist as yet

Mark 1 Geographical Analysis Machine

- **an early attempt at automated exploratory spatial data analysis that was easy to understand**
- **it answered a *simple practical question***

  given some X,Y point referenced data of something interesting WHERE might there be evidence of localized clustering if you do not have the foggiest idea of where to look due to lack of knowledge?

Geographical Analysis Machine (GAM) Mark 1 history

- GAM/1 developed in the mid 1980s
- it was very computationally intensive hence the term **MACHINE** because it really needed a dedicated computing machine
- Early runs took over 1 month of CPU time on a large Mainframe (Amdahl 580)
- Later ran on a Cray X-MP, Y-MP, and Cray 2 super computers

- It was developed to analyze Child Leukemia Data in Northern England
- GAM/1 easily spotted the suspected Sellafield Cluster
- **BUT**

  it also found an even stronger major new cluster in Gateshead Cancer Cluster in 1986

- This is possibly the **ONLY** instance of a major cancer cluster being found by analysis (rather than journalism) since John Snow's famous cholera spatial epidemiology of the mid 19th century

10 years ago GAM/1 was a mixed blessing!

- It was praised by many geographers as a major development in useful spatial analysis technology
- It was severely criticized by some statisticians (mainly due to ignorance of the geography of the problem)
- Software for GAM was never distributed as ten years ago it was not easily run

GAM/1: good aspects

- it was automated
- prior knowledge or ignorance was rendered irrelevant
- it looked for localized clusters at a time when most spatial statistical methods concentrated on global measures of pattern
- the search for local clusters was geographically comprehensive

- all locations were treated equally
- spatial data imprecision was explicitly handled (a major first)
- the results were study region boundary invariant
- the outputs were cartographic rather than expressed in terms of complex statistics
- it suggested hypotheses that could be tested by other methods later
- it was an early example of a geographical data mining tool
- it was suggested as a prototype of a whole class of equivalent methods

GAM/1: Bads

- it needed a supercomputer and was not easy to apply because of restricted access
- there are multiple testing problems
- it upset some major statisticians who conduct a brief campaign of intensive criticism most of which turned out to be either **incorrect** or **irrelevant** or **mischievous** in intent

- **Some of the causes for the dislike of GAM :**
  - the deliberate attempt at automating the artistic science of statistical analysis was disliked
  - the use of vast amounts of computer time wrongly thought to be an expensive rare commodity
  - the idea of localized clustering was initially but incorrectly considered to be implausible
  - lack of any well developed statistical theory
  - a preference for a Bayesian mapping approach
  - failure to consider data error as a cause of clustering
  - a prejudice against data mining
  - media hype, exaggeration, and high public profile that surrounded the initial results

and

it was developed by a

GEOGRAPHER

who foolishly used the word STATISTICAL analysis rather than the term GEOGRAPHICAL ANALYSIS too frequently!

Some of the problems went away

- Criticisms faded away in the early 1990s as spatial statisticians developed a better understanding of the geography of the problem and the statistical concerns were better understood
- Besag and Newell developed their “own” version of GAM
- The Gateshead results were subsequently corroborated but their cause remains a mystery

International Agency for Research on Cancer (IARC)

- Commissioned a study in 1989-91 of different clustering methods, many developed by critics of GAM
- 50 synthetic cancer data sets were created for which the degree of clustering and locations of clusters were known but kept secret
- the data were given to the participants who performed their analyses without any knowledge of the correct results

- **Methods applied <u>without knowledge</u> of the results were**
  - Potthoff-Whittingham
  - Cuzick-Edwards
  - GAM-K
  - Besag-Newell's method
  - ISD's Original Method
- **Later extended to include 4 others but these were applied <u>with knowledge</u> of the cluster locations**
  - ISD revised
  - Cuzick-Edwards one sample method
  - Diggle-Morris K functions
  - CAS method

Results published in Alexander and Boyle (1996)

- Expected publication date was 1991
- It was anticipated that the statistical methods preferred by the critics of GAM would work best
- Much to the **SURPRISE** of Alexander and Boyle GAM/K was shown to be the best or equivalent best means of **TESTING FOR PRESENCE OF CLUSTERING and for FINDING THE LOCATIONS OF CLUSTERS**

An inverse relationship between the strength of GAM criticism and the performance of the preferred methods!

- **The widely favoured Bayesian Mapping was dropped altogether as it was highly misleading!**
- **The ISD's Methods were poor and unreliable clustering detectors and were not designed to find clusters**
- **Besag-Newell's modified GAM was untrustworthy with a high false alarm rate**
- **Diggle's much vaunted K Functions were next to useless for non-random data!**

Clustering Found in 10 Random Data Sets

Potthoff-Whittingham	**2**
Cuzick-Edwards	**2**
GAM-K	**2**
Besag-Newell	**8**
ISD original	**5**
ISD revised	*4*
Cuzick-Edwards	*1*
Diggle-Morris	*0*
CAS	*4*

Clustering Found in 13 One Cluster Data Sets

Potthoff-Whittingham	**2-6**
Cuzick-Edwards	**7**
GAM-K	**11**
Besag-Newell	**8**
ISD original	**8**
ISD revised	*8*
Cuzick-Edwards	*8*
Diggle-Morris	*2*
CAS	*11*

Clustering Found in other clustered Data Sets

Potthoff-Whittingham	**1-8**	**4**	**6**
Cuzick-Edwards	**9**	**4**	**6**
GAM-K	**11**	**5**	**8**
Besag-Newell	**9**	**4**	**7**
ISD original	**10**	**5**	**7**
ISD revised	*10*	*5*	*7*
Cuzick-Edwards	*10*	*4*	*7*
Diggle-Morris	*5*	*4*	*2*
CAS	*12*	*5*	*8*
Numbers of Clusters	**12**	**7**	**8**

Detection of Cluster Locations

- A much harder problem
- Cancer is a RARE event and the synthetic data contained considerable amounts of random noise
- Risk Factors were 2D Gaussian with different spatial variances and levels of clustering
- Not all the apparent cluster locations were findable!!

Finding Correct Clusters

Findable Sites Found “near” real parent locations

	1 parent	**2**	**8**
Besag-Newell	10	20	17
GAM/K	10	16	12
Cuzick-Edwards	7	10	16

Finding False Clusters

Sites Found "near" parent locations

	Random	**1**	**2**	**8**
Besag-Newell	8	10	5	3
GAM/K	2	3	1	0
Cuzick-Edwards	2	2	3	0
Totals	**10**	**13**	**12**	**7**

Estimated Positive Sensitivities

Besag-Newell	**36 %**
Cuzick-Edwards	**66 %**
GAM/K	**87 %**

Alexander and Boyle (1996) authors of the IARC study concluded:

"The GAM has potential applications in this area if adequate computer resources are available. At the present time, however, the new, more sophisticated version of the GAM is complex, difficult to understand.."

(p 157)

That was in *1991*!!!!!!!!

- **There were THEN TWO remaining criticisms**:
  - (1) GAM needed a supercomputer
  - (2) GAM was complex
- **Others could have been added**
  - (3) GAM was not available for others to use
  - (4)GAM linkages with GIS was unclear
- **Are these criticisms still valid today?**

Reviving GAM/K

- GAM/K still runs on the later day version of the Cray X-MP vector supercomputer (the Cray J90)
- Efforts were made in 1996-7 to port the Cray X-MP code on to a Cray T3D parallel supercomputer with 512 processors
- **BUT it failed!!!**

Original GAM/K Cray X-MP code

- **Code was designed to work with UK level postcoded cancer data (100m resolution)**
  - it could not handle local crime data based on Address Point grid-references (1m resolution)
  - it was no longer a general purpose method
- **The code assumed a maximum memory of 1 megaword (a lot in 1986)**
  - it used data compression methods to reduce memory sizes but this is no longer necessary and added considerable complexity

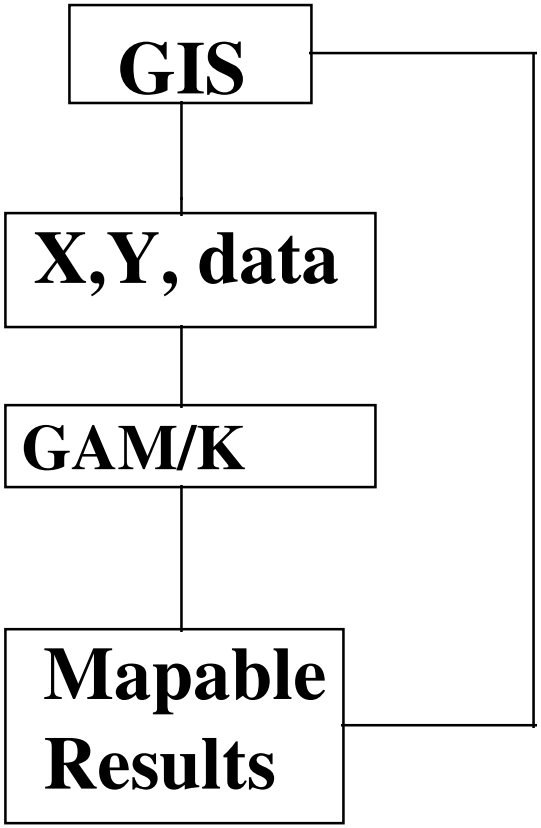
Algorithm was re-programmed from scratch

- **But** it needed an estimated 9 Days of CPU time on a single J90 processor (or Cray T3D node) to perform a single UK run using census Ed data
- ***Fortunately:***
  - a modern workstation is as fast as a Cray J90 processor
  - a single run could be completed in less than 1 hour on a 512 processor Cray T3D, however..
- **<u>This would hardly constitute a generally applicable and easy to use method!</u>**

Making GAM/K run faster

- **Subsequent modifications to the spatial data retrieval algorithm used in GAM/K reduced the 9 days to 714 seconds on a workstation**
- Total M/flops reduced from 10,498 million to 46 million
- **GAM/K was now a PRACTICAL tool**
- It can be readily linked to any GIS once it no longer needed a supercomputer to run it

Linking GAM/K to any GIS is easy



How does GAM/K work?

- **Two Parts**
  - a cluster search part
  - a pattern recognition part
- **Part 1:** is the statistical part that repeatedly applies some test of localized clustering to all parts of the map
- **Part 2:** attempts to integrate all the available evidence to identify any patterns that appear to matter from those that do not

Basic GAM/K algorithm: Part 1

- do circle_RADIUS=min_RADIUS, max_RADIUS
- do circle_X=min_X, max_X, circle_RADIUS*0.2
- do circle_Y=min_Y, max_Y, circle_RADIUS*0.2
  - get data for circle at (circle_X, circle_Y, circle_RADIUS)
  - assess its significance
  - if significant then write (circle_X, circle_Y, circle_RADIUS)
- enddo
- enddo
- enddo

Basic GAM/K algorithm: Part 2

- initialize a raster map grid of 1000 by 1000
- read circle_X, circle_Y, circle_RADIUS
- compute “excess” (observed-expected cases)
- apply Epanechnikov kernel to create density estimated using circle_RADIUS as bandwidth
- surface summed over raster over map grid
- repeat for all “significant” circles
- map surface
- apply Cluster detector or eye-ball!

Get Data for Circle (CX,CY,radius)

- Naive algorithm

```
POP=0
CASES=0
do i=1,n
 dis=sqrt((x(i)-CX)**2+(y(i)-CY)**2+0.001)
   if(dis.LE.radius)then
          POP=POP+p(i)
          CASES=CASES+c(i)
          endif
enddo
```

- If radii=1 to 50, UK map area, and n=145 000 then

  3, 335, 000, 000, 000

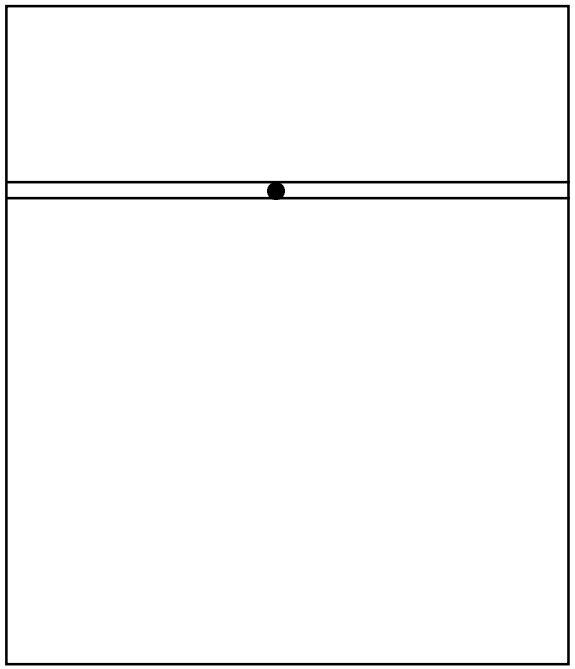
  DO LOOP executions = 8 days!

A more efficient circle retrieval algorithm

- Can reduce compute times by a factor of 10,000 or more

0.001^{th} to 0.01^{th}

of the data



Burglary Data for Sheffield

- Part of Sheffield analyzed
  - 71,911 Address Point houses
  - 3,784 burglaries
- Comparative results available from STAC system
- All of South Yorks could have been studied just as easily and almost as rapidly!

Crime Data is considered to be far easier to analyse that rare disease data

- crime incidence is not rare
- little or no latency to weaken patterns
- high quality spatial data
- results have no or low media value
- spatial-statistical problems should be greatly reduced

What significance test?

- Method used depends on rarity of incidence of interest
- If “leukemia rare” then Poisson probabilities are appropriate
- If “fairly common” then Poisson or binomial may be appropriate
- If “very common” then binomial or bootstrap an excess statistic may be appropriate
- If in doubt then use a Monte Carlo Significance test

Results: uses Poisson Probabilities

Circle size	Sites Examined	Significant Circles	Time secs
200m	54,450	2,307	15.7
400m	14,352	1,137	11.1
600m	6,716	523	9.6
800m	3,933	242	8.7
1km	2,688	155	8.3
2km	700	22	7.0
3km	323	6	6.3
4km	182	0	5.6

Machine used?

- ☒ **Cray J90 vector supercomputer?**
- ☒ **Cray T3D 512 processor parallel supercomputer?**
- ☑ **Sun 10/41 5 year old unix workstation?**

Results: uses Poisson Probabilities with Random Data

Circle size	Sites Examined	Significant Circles	Time secs
200m	54,450	63	14.9
400m	14,352	12	10.6
600m	6,716	5	9.4
800m	3,933	3	8.6
1km	2,688	0	8.2
2km	700	0	7.0
3km	323	0	6.3
4km	182	0	5.6

Results: uses Binomial Probabilities

Circle size	Sites Examined	Significant Circles	Time secs
200m	54,450	1,946	4.1
400m	14,352	1,028	2.8
600m	6,716	458	2.1
800m	3,933	198	1.8
1km	2,688	129	1.7
2km	700	16	1.4
3km	323	3	1.3
4km	182	0	1.1

Results: uses Monte Carlo Test

Circle size	Sites Examined	Significant Circles	Time secs
200m	54,450	2,315	2837
400m	14,352	1,149	1745
600m	6,716	716	945
800m	3,933	524	722
1km	2,688	397	472
2km	700	143	254
3km	323	77	191
4km	182	47	108

What about multiple testing?

- **If alpha=0.05 then**
  - if you test 100 hypotheses then 5 will be false positives (i.e. appear as significant wrongly)
  - if you test 1,000,000 hypotheses then 50,000 will appear as false positives

<u>IF the tests are INDEPENDENT</u>

- *GAM hypotheses are not independent hence need to use Monte Carlo Test to check on this aspect or ignore it!*

Multiple Testing Results

Circle size	Sites Examined	Significant Circles	probability random
200m	54,450	2,307	0.009
400m	14,352	1,137	0.009
600m	6,716	523	0.009
800m	3,933	242	0.009
1km	2,688	155	0.009
2km	700	22	0.009
3km	323	6	0.006
4km	182	0	1.0

Multiple Testing Results: for Random Data

Circle size	Sites Examined	Significant Circles	Probability random
200m	54,450	63	.594
400m	14,352	12	.782
600m	6,716	5	.722
800m	3,933	3	.623
1km	2,688	0	1.0
2km	700	0	1.0
3km	323	0	1.0
4km	182	0	1.0

Conclusions

- **Results**
- **Other GAMs**
- **Future Plans**

GAM Strengths

- Totally automated analysis
- Can find multiple clusters in a single run
- Results are largely self-evident
- Method is very easy to explain in plain English without statistical complexity
- Fast

GAM Deficiencies

- Brute Force search
  - dumb
  - does not scale to higher dimensional spaces
- Compute Intensive
- Mapped results may mislead
- Not available in GIS yet
- Animation potential is poor

The other GAMs

- MAP EXplorer (MAPEX) is an intelligent search version of GAM/K
  - uses Genetic Algorithm to perform search
  - uses AVS to create MPEG computer movies of search process
- Space Time Attribute Creature (STAC) extends MAPEX to multiple data domains
  - uses Java and web browser for animation
  - interactive

Future Plans

- A strategy of distributing GIS independent portable code to maximize usefulness
  - versions planned for Arc/Info and Map Info
- Development of a Java visualization animation add-on
  - later 1997
- ESRC project to Implement GAMs as an Internet form of distributed geographical analysis over next 2 years

Further Info: Email

stan@geog.leeds.ac.uk