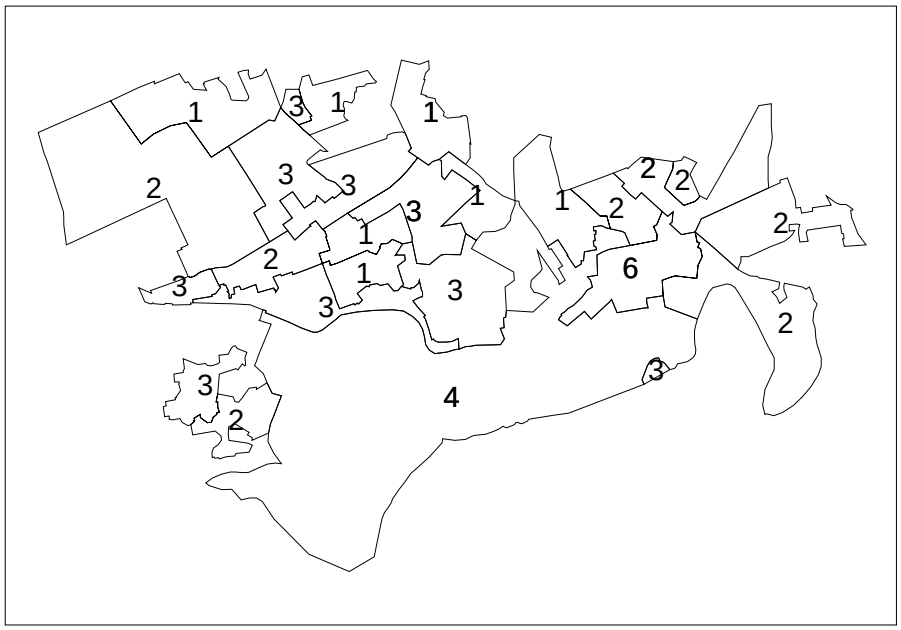
# Motivating Examples

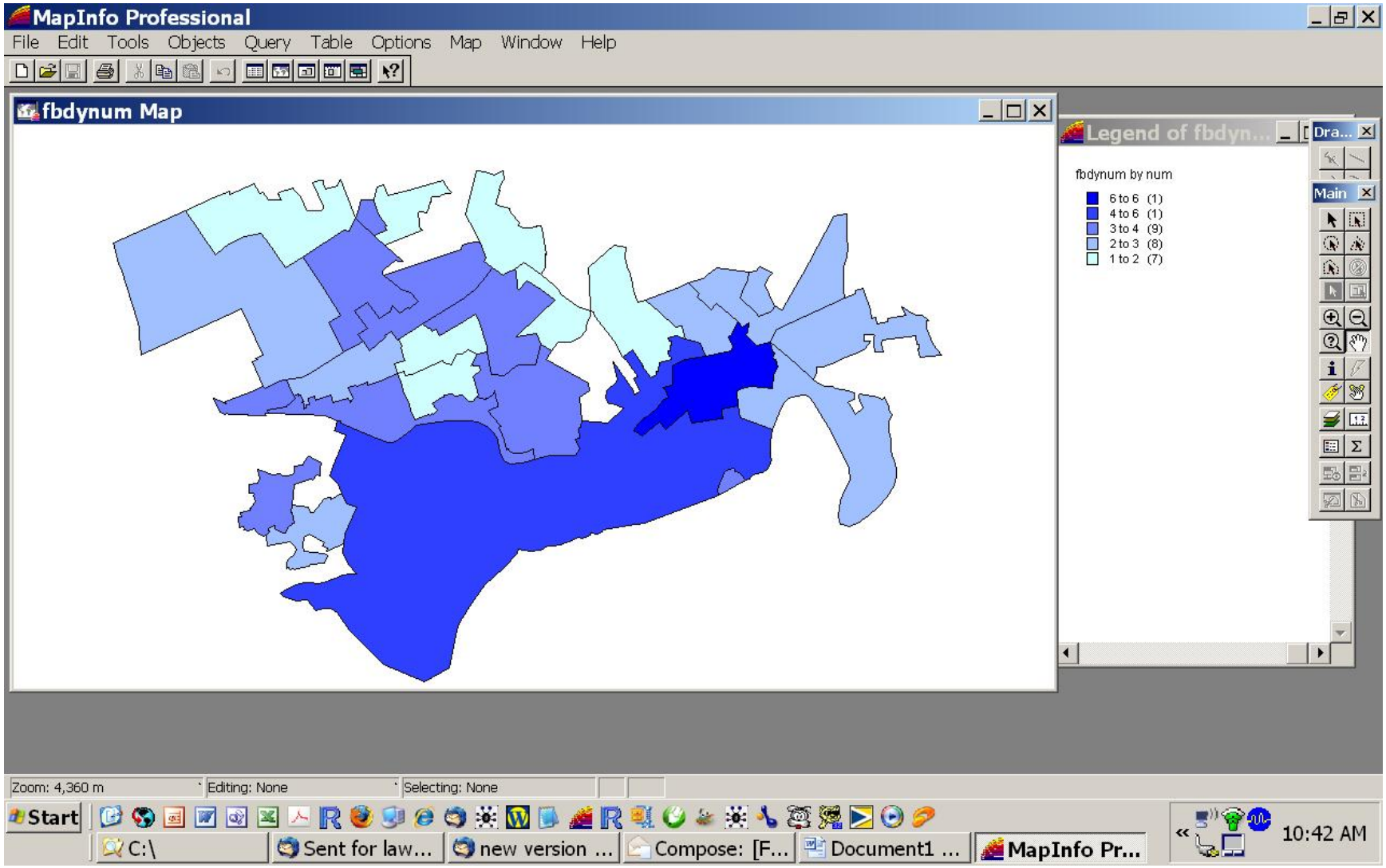
- Falkirk: East central Scotland (UK)

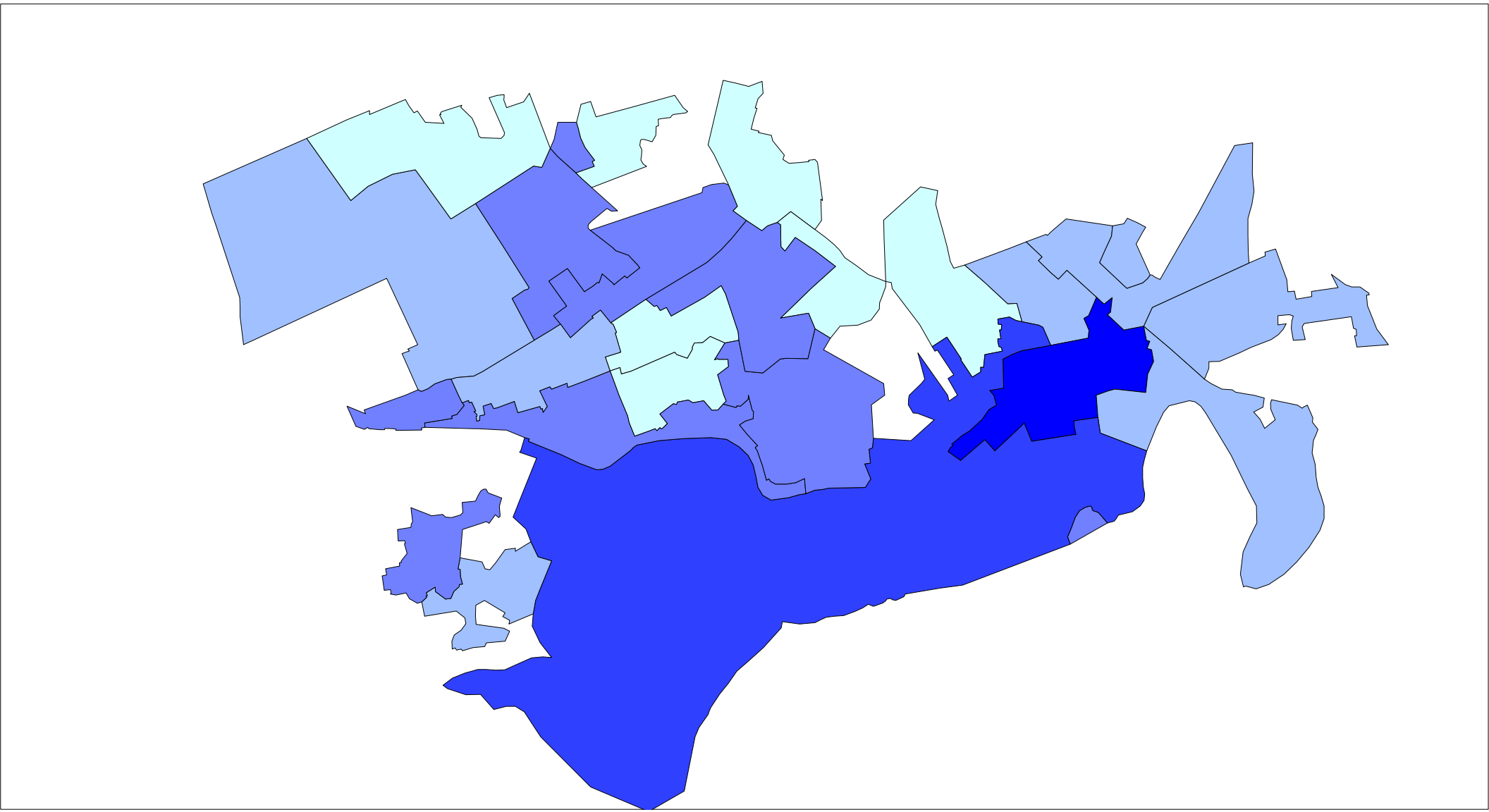
- crude counts of respiratory cancer (ICD code 162)

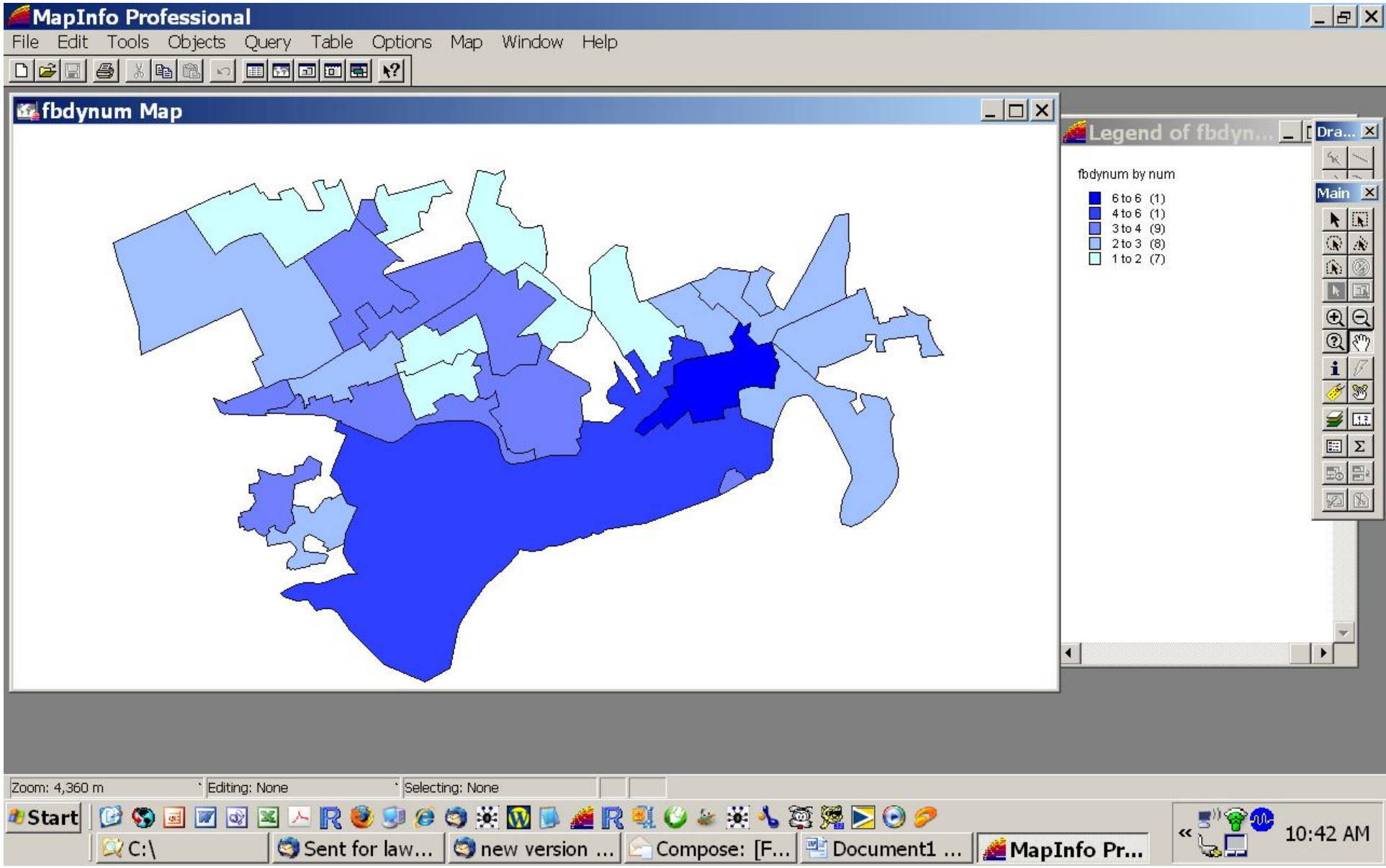
- five year time period (1978-1983)

- 26 enumeration districts (census tracts)

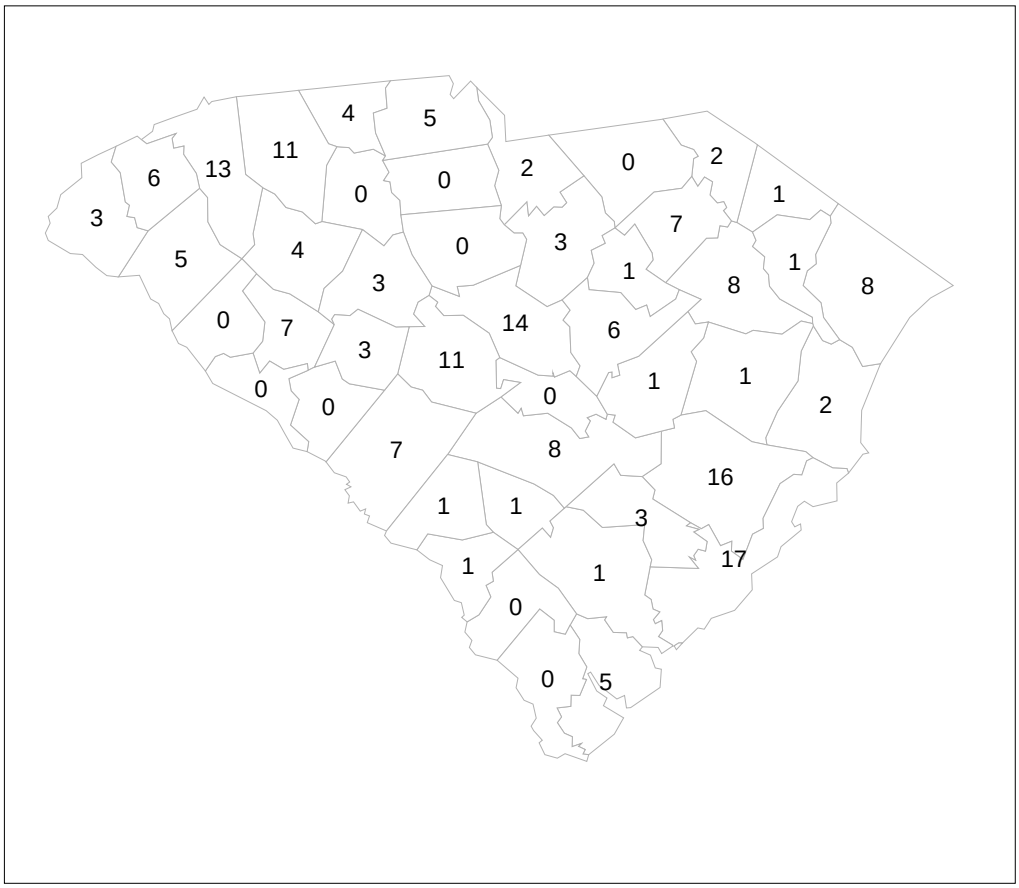

1
3
1
1
2
3
3
1
3
1
1
2
2
2
2
6
2
1
3
3
3
2
3
4
3
2


MapInfo Professional
File Edit Tools Objects Query Table Options Map Window Help
fbdynum Map
Legend of fbdyn...
fbdynum by num
6 to 6 (1)
4 to 6 (1)
3 to 4 (9)
2 to 3 (8)
1 to 2 (7)
Dra...
Main
Zoom: 4,360 m
Editing: None
Selecting: None
Start
C:\
Sent for law...
new version ...
Compose: [F...
Document1 ...
MapInfo Pr...
10:42 AM




MapInfo Professional
File Edit Tools Objects Query Table Options Map Window Help
fbdynum Map
Legend of fbdyn...
fbdynum by num
6 to 6 (1)
4 to 6 (1)
3 to 4 (9)
2 to 3 (8)
1 to 2 (7)
Dra...
Main
Zoom: 4,360 m
Editing: None
Selecting: None
Start
C:\
Sent for law...
new version ...
Compose: [F...
Document1 ...
MapInfo Pr...
10:42 AM

- South Carolina:
- total congenital abnormality deaths
- 1990 by county



# Fundamental Features

- Case events occur at residential addresses and the count of cases within regions is made
- Regions are of arbitrary extent and geometry (often electoral districts or postal zones)
- Cases arise within a heterogeneous population, which varies in space and in nature
- The residential location of the cases can also be analysed
- The purpose of mapping can be :

- - to assess spatial variation of disease to assess Public Health needs

  - to assess whether unusual clusters/concentrations of disease are evident

  - to examine connections between disease incidence and environmental or other explanatory covariates which have spatial expression

# Basic Epidemiological Concepts

- Relative Risk

- crude counts of disease cannot be used on their own: why?

- need to examine the population from which the counts arose

- the background ‘at risk’ population varies in its density and composition with spatial location

- often we want to compare the observed count with what would have arisen from the underlying population

- often a ratio of the observed count $\{n_i\}$, in the *i th* tract to the expected count $\{e_i\}$ derived from the background population:

relative risk as a ratio: $\frac{n_i}{e_i}$ in any region/tract

- what other ways can we compare the observed to expected rates?

# Standardised Mortality/Morbidity Ratio(SMR)

- Mortality: disease recorded is a death

- Morbidity: disease recorded is a live case
- Incidence: occurence of the disease within a (usually short) time period. This is sometimes used to imply morbidity (live cases)
- Prevalence: as for incidence but over longer time period

- Standardisation: expected rates $\{e_i\}$ are calculated (estimated) from the local population structure i.e. the numbers in each tract in different age × sex groups are multiplied with known rates for the disease for equivalent groups in a *standard* population.

- The standard population may be the *national* population or even the *study region* population. Once the rate for each group is estimated the rates are added to give a total expected rate for the tract $\{e_i\}$.

- The standardised ratio of either mortality or morbidity is the relative risk ratio computed with standardised expected rates, as specified above:

$$SMR = \frac{n_i}{e_i}$$


Falkirk by smr
1.26 to 2.04 (6)
1.06 to 1.26 (3)
0.8 to 1.06 (6)
0.61 to 0.8 (5)
0.3 to 0.61 (6)

# Some notation

- For each of the *m* regions on the map:
- $y_i$ or $n_i$ the count of disease in the $i^{th}$ region
- $e_i$ is the expected count in the $i^{th}$ region
- $\theta_i$ is the relative risk in the $i^{th}$ region
- The SMR is just $smr_i = y_i / e_i$
- This is just an estimate of $\theta_i$

# Standardization

- Direct: conversion of counts to rates
- Indirect: compute the expected number of cases based on a reference population
- Indirect used here extensively as we have rare diseases and sparse data
- Computation: must know reference population rate
  - e.g. counties and statewide rate ($R$):

$$R = \sum_{counties} y_i \Big/ \sum_{counties} p_i, \text{ where } p_i \text{ is the county population}$$

  - Then:

$$e_i = p_i.R = y_T.(p_i / p_T)$$

# SMR problems

- Notoriously unstable
- Small expected count can lead to large SMRs
- Zero counts aren't differentiated
- The SMR is ***just the data!***

# Smoothing for risk estimation

- Modern approaches to relative risk estimation rely on smoothing methods
- These methods often involve additional assumptions or model components
- Here we will examine only one approach: ***Bayesian modeling***

# Control Diseases and Expected Rates

- Expected rates are commonly used to allow for population effects
- An alternative is to use the incidence of a *control disease* within each region/tract
- A control disease is matched closely to the risk structure for the disease of interest, but must not display the incidence effect under investigation
- Examples:
  - coronary heart disease is often used as a control disease for respiratory cancer in studies of air pollution
  - live birth numbers could be used as a control for childhood leukaemia in clustering studies
  - lower body cancers (testes, penis, ovaries) can be used for respiratory tract diseases
- The common feature of each control is that it shouldn't be related to the effect of interest
- There is much debate about use of these controls as opposed to expected rates from external sources.

# The Ecological and Atomistic Fallacy

- Many mapping studies attempt to relate incidence of disease in regions with some other measurable *explanatory* variable relating to the aetiology of the disease e.g. we might want to examine the relation between the number of smokers in regions and the incidence of respiratory cancer in the same regions. This might be achieved by applying regression analysis to the disease incidence and explanatory variable.

- The relation between these variables will be statistical and may suffer from the fact that regional totals or averages are used in the assessment of the relation. Hence an average relationship can only be measured

- The *ecological fallacy* arises when such regional average characteristics are ascribed to *individuals* within the region concerned Any region-based analysis will suffer from this problem

- The *atomistic fallacy* occurs when analysis is based on individuals, and the variability of individuals' response to disease is not accounted for in inference at the regional level.

# Confounders and Deprivation Indices

- All disease maps contain the influence of variables affecting, or pertaining to, the local population which are not accounted for in standardised rates or control diseases.

- We can try to allow for these effects in two ways:

- - Include as many known explanatory variables in the expected rate to allow for these effects . These variables are called known *confounders.*

- - Include the effect of unknown confounders via the use of *random effects.*

- Often adverse disease incidence is related to a range of poverty-related explanatory variables e.g. unemployment, housing type,welfare status, car ownership. these variables are often available from national census. There has been some effort recently to combine such variables in composite measures known as *deprivation indices.*

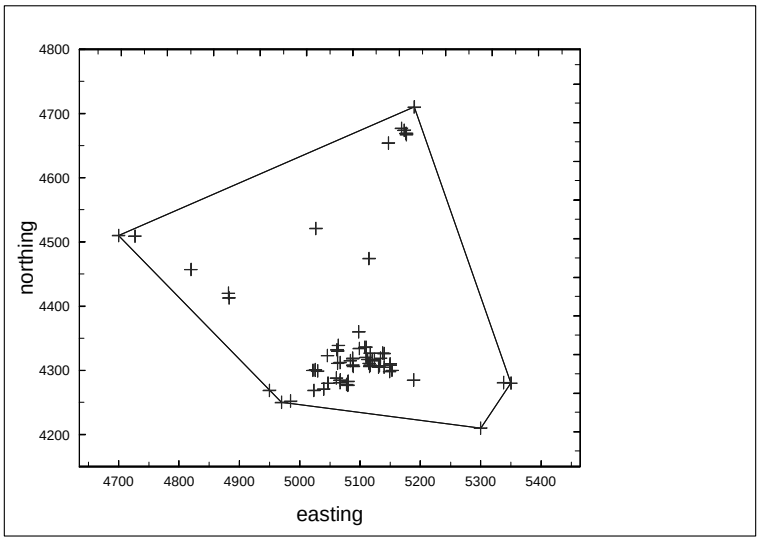
- Deprivation indices are now routinely available from government census data organisations and can be incorporated directly into a disease map

# Data types and exploratory analysis

Two basic data formats arise in disease mapping: case event and tract count

Case event data: the *locations* of cases (usually residential location) are the basic observational unit, they

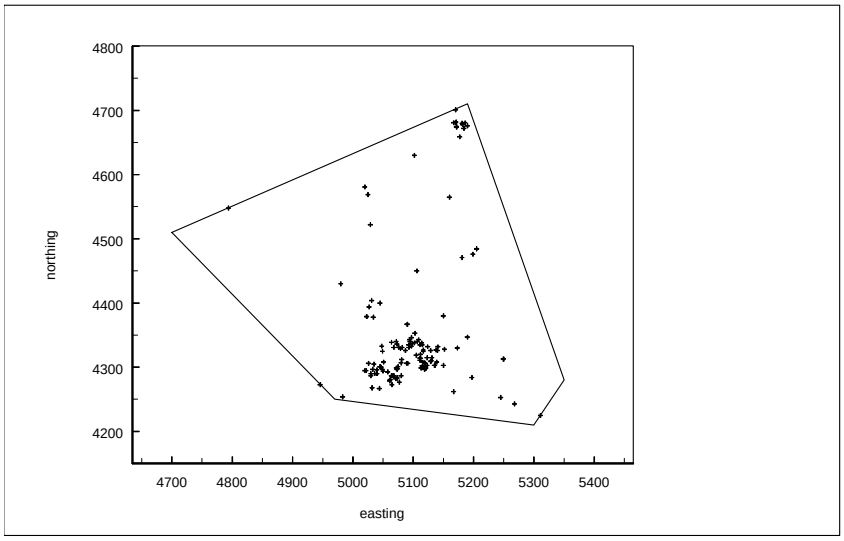
- could be street address or grid reference




case event data requires analysis based on point process methodology i.e. the locations are a sample from a density governing the spatial distribution of case events. Further analysis requires the introduction of a *first order intensity function* which describes the spatial distribution of the cases. This intensity function can be normalised to yield a density.

As for count data, we need to estimate or make allowance for the population locally which gave rise to the events. Often, due to the disaggregation of the events, a sample of a control disease is used for this purpose.

The first order intensity of the control disease is then estimated to provide an *expected* intensity of cases




Tract count data: the number of case events recorded within the tract of interest

Only a count is held and the exact locations of cases is unknown, although the tract centroid may be known

The local disease incidence can be expressed directly as the count of disease in a tract, while externally standardised rates can be used to make allowance for the local population structure.

# Exploratory Analysis of Maps

Usually the relative risk surface over the study region is examined to yield information about local concentrations of disease (clusters) or trends in the disease incidence

- Case event data:

- form the ratio $\{R(\mathbf{x})\}$ of the first order intensity of cases $\{\lambda(\mathbf{x})\}$ to the first order intensity of the control sample/realisation $\{g(\mathbf{x})\}$, where $\mathbf{x}$ is a spatial location:

$$\widehat{R(\mathbf{x})} = \frac{\widehat{\lambda(\mathbf{x})}}{\widehat{g(\mathbf{x})}}$$

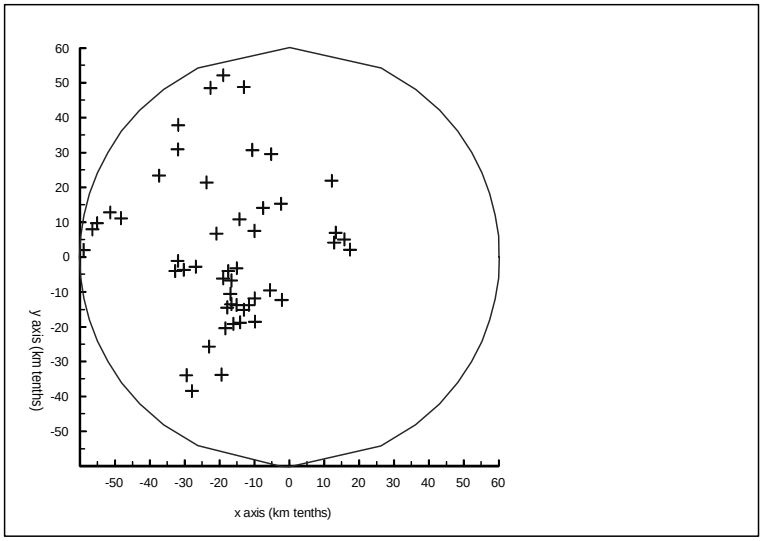
Both intensities can be estimated separately by two dimensional density estimation

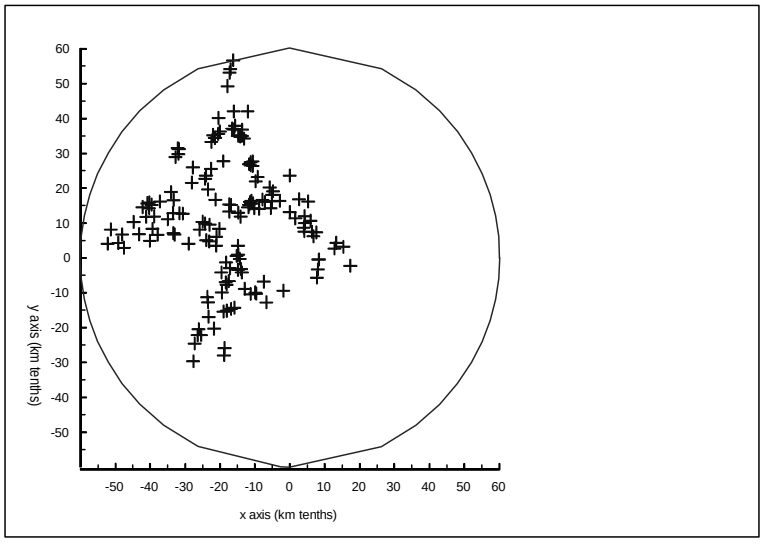
This ratio is an estimate of relative risk across the study region.

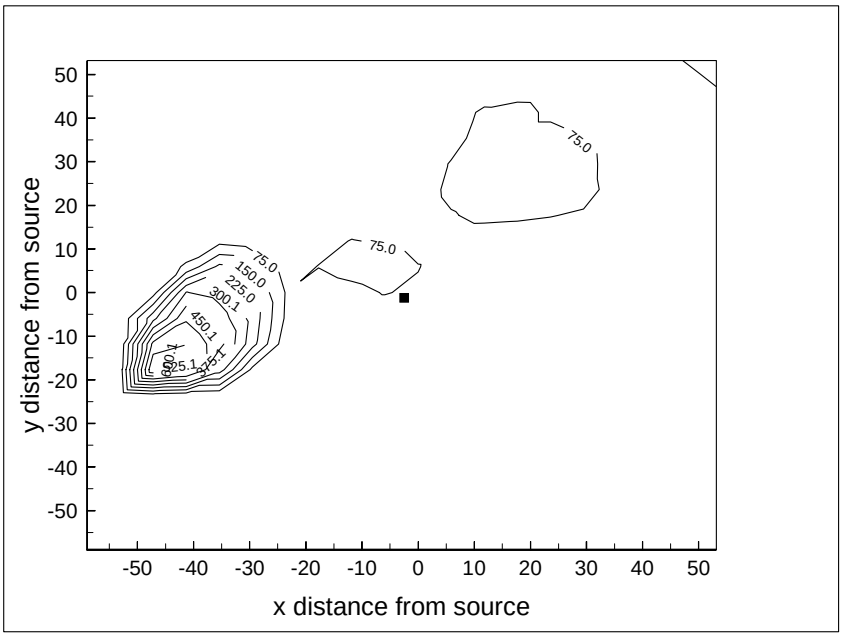
- Count data:

- form the ratio of observed to expected counts in each tract and this SMR estimate gives a tract-wise constant relative risk estimate. This can be converted to a continuous relative risk surface via smoothing methods (e.g. kernel regression):

$$\widehat{R(\mathbf{x})} = S\left\{\frac{n_i}{e_i}\right\}$$


60
50
40
30
20
10
0
-10
-20
-30
-40
-50
y axis (km tenths)
-50 -40 -30 -20 -10 0 10 20 30 40 50 60
x axis (km tenths)


60
50
40
30
20
10
0
-10
-20
-30
-40
-50
y axis (km tenths)
-50 -40 -30 -20 -10 0 10 20 30 40 50 60
x axis (km tenths)


y distance from source
x distance from source
50
40
30
20
10
0
-10
-20
-30
-40
-50
-50
-40
-30
-20
-10
0
10
20
30
40
50
75.0
75.0
75.0
150.0
225.0
300.1
450.1
600.1
675.1
750.1