
KAREN KAFADAR,* LAURENCE S FREEDMAN,** COLIN R GOODALL† AND JOHN W TUKEY‡


Background. The effect of urbanization on age-adjusted lung cancer mortality rates in US counties is investigated. The data come from National Cancer Institute, and urban trends are estimated in time periods 1970–1979 and 1980–1987, for both white males and white females. To account for possibly different gradients in different parts of the country, the 48 contiguous states are divided into seven regions.

Methods. A measure of urbanness, urbanicity, is defined and is used to stratify counties. A multiplicative model is proposed that relates county mortality rates to urbanicity. The residuals from this multiplicative model serve as age- and urban-adjusted rates.

Results. Urban-rural gradients are significant for nearly all regions for both white males and white females, diminishing slightly in the latter time period for white males but becoming stronger for white females.

Conclusions. The age- and urban-adjusted rates may be used in mapping to investigate geographical patterns that remain after removal of the urban factor.

Keywords: geographical epidemiology; urbanicity; smoothing; Poisson; exploratory data analysis

In geographical analysis of incidence or mortality rates, it is useful to map the disease rates and then identify those areas for which the rates are 'statistically high' (or low) according to some assumed distribution for the numbers of deaths (usually Poisson). These rates are usually adjusted for age. Adjustment for other known risk factors would help in eliciting further underlying geographical patterns (beyond those described by the risk factors).

Rate adjustment for known risk factors is often performed for age but less often for other factors. A different type of adjustment, namely for variance inhomogeneity, has been applied, for example, by Clayton and Kaldor, Manton et al. and Cressie. They use empirical Bayes methods to shrink rates based on few person-years of observation towards the mean rate. We will not discuss this type of adjustment in this paper.

In this paper we describe a geographical analysis of lung cancer mortality rates in US counties which is preceded by adjustment of the rates not only for age but also for urbanicity. We choose urbanization as a variable for adjustment ("adjustor") because urbanization correlates highly with smoking and may also correlate with other risk factors, such as exposure to environmental carcinogens, individual lifestyles, air quality, ethnic composition of the population, manufacturing plants, and reporting of diagnostic procedures. We cannot adjust for smoking directly because detailed data on smoking behaviour is available for only some counties, and surveys such as NHANES by the National Center for Health Statistics and the Behavioral Risk Factors Survey by the Centers for Disease Control provide smoking data only within broadly defined regions of the US. Since we have county mortality rates, such data are not fine enough for our purposes. For these two reasons, correlation with other risk factors besides smoking and availability of data, we show that a carefully
defined measure of urbanization serves very well as an adjustor.

This study continues the analysis in Kafadar and Tukey² on lung cancer mortality in white males for 1950–1969. Here we consider the mortality rates for cancer of the lung, trachea, bronchus, and pleura, for two later periods, 1970–1979 and 1980–1987, for two groups, white males and white females. These sites are selected for three reasons. First, among all cancer sites, lung cancer mortality rates were highest for males, and second highest for females, in the periods studied, offering a strong ‘signal’ for analysis. Second, urbanicity is an obvious variable for adjustment (Section 3) and illustrates clearly the methodology. Third, the accuracy of lung cancer as cause of death on death certificates is presumed to be quite good.⁶ The mortality rates have been obtained from the National Cancer Institute (NCI) and are based on death certificates which include legal US residence at time of death.⁷ The total numbers of deaths during 1970–1979 and 1980–1987 for the continental US have been age-adjusted using the sum of the mid-year US census population estimates relative to the 1970 US population.

Section 2 describes previous work on the effect of urbanization on US lung cancer mortality. Section 3 describes the methods of analysis: subdivision into regions and stratification by urbanization. The parameters from a multiplicative model relating mortality to a measure of urbanization are estimated and serve as the basis for an urban adjustment of the rates, the results of which are given in Section 4 and discussed in Section 5. Final comments and proposals for further work are given in the closing section.

THE ASSOCIATION BETWEEN URBANIZATION AND LUNG CANCER

Several authors have noted an urban-rural gradient with respect to lung cancer mortality rates. Excess mortality in urbanized areas was noted in atlases of US cancer mortality among both whites (ref.⁸ p. 15) and non-whites (ref.⁹ p. 16). Shy¹⁰ summarizes the results of several studies, many of which show a significant urban-rural gradient when urbanization is categorized according to the population of the unit (e.g. town, county). Four of these studies are prospective and indicate that the urban-rural difference is larger among non-smokers than among smokers, suggesting that other components of urbanization are important. Haenszel et al.¹¹,¹² use data from the 1958 Current Population Survey and show that higher (lower) cancer risk was noted for movement from smaller to larger places (larger to smaller places).

Most analyses of the urban effect on lung cancer mortality (or other sites) measure urbanization as total population of the area, and then characterize its effect as the relative mortality ratio in two, three, or four urban categories.¹⁴,¹⁵ Dichotomous classifications according to per cent urban, as defined by the US Bureau of the Census, have also been considered.¹⁶ Nasca et al.¹⁷ calculated average lung cancer incidence in New York counties stratified by ranges of population density. Doll¹⁸ reported on a comparison of lung cancer mortality between 13 ‘wholly urban’ counties and 957 ‘wholly rural’ counties among the 3000+ US counties. In all these studies, a ‘significant’ ratio between ‘urban’ and ‘rural’ counties was indicated. Thus it seems sensible to adjust rates of lung cancer mortality for the urban trend before seeking out geographical patterns other than urban/rural ones which may be present.

It should be emphasized that adjustment for urbanicity is only in part a proxy for smoking prevalence within the county. In fact, the strength of the correlation between smoking prevalence and urbanization appears to be weakening over the past four decades (Blot, personal communication). Nonetheless, there are probably other relevant components of the urban factor, such as environmental hazards, for which it will be important to adjust cancer mortality rates.

METHODS: STRATIFICATION BY REGION AND URBAN INDEX, AND A MODEL FOR ADJUSTMENT

Regions for analysis

The notion of urbanization may well differ depending upon the region of the county. For example, the presence of even one very large city in Laramie county, Wyoming (Cheyenne, 1980 population = 58 265) may influence the perception of urbanization among its residents very differently than it would for residents of Morris county, New Jersey (largest place is Parsippany, 1980 population 49 868). (The next two largest places in Laramie have populations of 5310 and 2767; in Morris, the next two largest places have populations of 19 850 and 18 878.) Moreover, when adjusting mortality rates prior to examination of geographical patterns, the adjustment may be more precise if it is based on a local rather than a national analysis. For these reasons, the country is subdivided into regions which are moderately segregated with respect to the urban index. With few exceptions, these regions turned out to be very similar to the US Census-defined regions¹⁹ and are shown in Figure 1. We believe these regions will be useful in examining the urban effect on mortality rates of other cancers.
A measure of urbanization: Urbanicity
To describe the characteristics of an urban place, one naturally thinks of big cities. The presence of even one major metropolitan area can have an impact on the inhabitants throughout the county. Kafadar and Tukey\(^5\) characterized the urbanness of a county in terms of the logarithm of the 1960 population of the largest urban place in the county, where urban place is defined and used by the US Bureau of the Census (ref.\(^19\) p. 4) and 1960 is the median of the years 1950–1969 covered in their data set.

The measure of urbanization used in this study is the county’s \textit{urbanicity},\(^20\) defined as ten times the base-2 logarithm of the root mean square of the three largest places in the county. Denoting these three populations from largest to smallest as \(p_1, p_2, p_3\), the urbanicity, \(U\), is defined by:

\[
U = 10 \log_2 \left( \frac{p_1^2 + p_2^2 + p_3^2}{3} \right)^{1/2}
\]

(1)

It is easy to see why this measure is highly correlated with a measure defined as ten times the base-2 logarithm of the total population in the county; in 2571 of the 3068 counties, the three largest places comprise over 50% of the county’s total population. Gross differences between the two measures occur primarily when the county’s population is spread more or less evenly throughout many more than just three places (e.g. certain suburban counties such as those on Long Island and in New Jersey). The logarithmic transformation conveniently symmetrizes the distribution of the measure: \(U\) ranges from 65 (Loving, Texas) to 216 (Los Angeles, California), with a mean of 135 and standard deviation of 19 for all 3068 counties in the continental US. The ranking of counties on this index seems sensible and faithful to one’s perception of urbanness. In several instances, both the major city and its nearest suburb(s) are in the same stratum (e.g. San Francisco with San Mateo and Alameda counties; Denver with Jefferson county; Washington DC with Montgomery county, Maryland). Further illustrations, details, and motivation for this measure and its relation to two other popular measures of urbanization, total population and population density, are given in Goodall \textit{et al.}\(^20\)

Since the two time periods under investigation are 1970–1979 and 1980–1987, all population counts are based on the 1980 census. For the most part, urban places are minor civil divisions (MCD) for eastern states and census civil divisions (CCD) for western states. Although MCD and CCD correspond only crudely to urban concentrations, the match between urban places and MCD/CCD is generally adequate for most counties. In some counties, fewer than three MCD/CCD could be identified; for such counties, the mean squared population involved however many places (1 or 2) that were given. Such instances are rare (48 ‘monads’ and 274 ‘dyads’) and are confined primarily to counties at either the very low end of the urban index scale (\(U < 105\)) or the very high end where the city itself defines the county (e.g. San Francisco, Baltimore City,
TABLE 1 Strata of the urban index \(U = 5 \log_2 \text{(mean squared population of three largest places)}\)

<table>
<thead>
<tr>
<th>Stratum</th>
<th>Urban index</th>
<th>Population*</th>
<th>No. of counties</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>65 &lt; (U) ≤ 100</td>
<td>90.5 &lt; (P) ≤ 1024.0</td>
<td>141</td>
</tr>
<tr>
<td>2</td>
<td>100 &lt; (U) ≤ 105</td>
<td>1024.0 &lt; (P) ≤ 1448.2</td>
<td>140</td>
</tr>
<tr>
<td>3</td>
<td>105 &lt; (U) ≤ 110</td>
<td>1448.2 &lt; (P) ≤ 2048.0</td>
<td>184</td>
</tr>
<tr>
<td>4</td>
<td>110 &lt; (U) ≤ 112.5</td>
<td>2048.0 &lt; (P) ≤ 2435.5</td>
<td>126</td>
</tr>
<tr>
<td>5</td>
<td>112.5 &lt; (U) ≤ 115</td>
<td>2435.5 &lt; (P) ≤ 2896.3</td>
<td>140</td>
</tr>
<tr>
<td>6</td>
<td>115 &lt; (U) ≤ 117.5</td>
<td>2896.3 &lt; (P) ≤ 3444.3</td>
<td>170</td>
</tr>
<tr>
<td>7</td>
<td>117.5 &lt; (U) ≤ 120</td>
<td>3444.3 &lt; (P) ≤ 4096.0</td>
<td>174</td>
</tr>
<tr>
<td>8</td>
<td>120 &lt; (U) ≤ 122.5</td>
<td>4096.0 &lt; (P) ≤ 4871.0</td>
<td>196</td>
</tr>
<tr>
<td>9</td>
<td>122.5 &lt; (U) ≤ 125</td>
<td>4871.0 &lt; (P) ≤ 5792.6</td>
<td>228</td>
</tr>
<tr>
<td>10</td>
<td>125 &lt; (U) ≤ 127.5</td>
<td>5792.6 &lt; (P) ≤ 6888.6</td>
<td>195</td>
</tr>
<tr>
<td>11</td>
<td>127.5 &lt; (U) ≤ 130</td>
<td>6888.6 &lt; (P) ≤ 8192.0</td>
<td>175</td>
</tr>
<tr>
<td>12</td>
<td>130 &lt; (U) ≤ 132.5</td>
<td>8192.0 &lt; (P) ≤ 9742.0</td>
<td>154</td>
</tr>
<tr>
<td>13</td>
<td>132.5 &lt; (U) ≤ 135</td>
<td>9742.0 &lt; (P) ≤ 11585.2</td>
<td>139</td>
</tr>
<tr>
<td>14</td>
<td>135 &lt; (U) ≤ 137.5</td>
<td>11585.2 &lt; (P) ≤ 13777.2</td>
<td>122</td>
</tr>
<tr>
<td>15</td>
<td>137.5 &lt; (U) ≤ 140</td>
<td>13777.2 &lt; (P) ≤ 16384.0</td>
<td>120</td>
</tr>
<tr>
<td>16</td>
<td>140 &lt; (U) ≤ 145</td>
<td>16384.0 &lt; (P) ≤ 21705.0</td>
<td>180</td>
</tr>
<tr>
<td>17</td>
<td>145 &lt; (U) ≤ 150</td>
<td>21705.0 &lt; (P) ≤ 32768.0</td>
<td>133</td>
</tr>
<tr>
<td>18</td>
<td>150 &lt; (U) ≤ 160</td>
<td>32768.0 &lt; (P) ≤ 65536.0</td>
<td>166</td>
</tr>
<tr>
<td>19</td>
<td>160 &lt; (U) ≤ 170</td>
<td>65536.0 &lt; (P) ≤ 131072.0</td>
<td>101</td>
</tr>
<tr>
<td>20</td>
<td>170 &lt; (U) ≤ 220</td>
<td>131072.0 &lt; (P) ≤ 4194304</td>
<td>84</td>
</tr>
</tbody>
</table>

* Corresponding interval on population scale if county had a single largest place of this size.

District of Columbia). Counties with no MCD/CCD (e.g. Yellowstone Park, Wyoming) were deleted from the analysis.

Because counties having roughly similar values of \(U\) are likely to be comparable in urban character, it seems sensible, for the purposes of adjusting mortality rates, to stratify the counties in the US according to a range of values of \(U\) (Table 1). This approach allows a more flexible adjustment than the traditional form based on linear regression of the mortality rates on the urbanicity measure. In this study, we fit a parameter for each stratum of urbanicity, smooth the resulting sequence, and then use the smoothed stratum estimates for adjustment. To achieve a fully effective adjustment, it is anticipated that one will need more than the 2–4 strata used in most previous studies. Since the precise nature of the urban trend is unknown, 20 strata of urbanicity offer sufficient stabilization after smoothing without sacrificing flexibility in the estimated trend.

Figure 2 gives a histogram of the values of \(U\) for the 3068 counties in the continental US. Table 2 gives the number of counties in each urban stratum in the seven regions, together with the average population during the years of these data (1970–1987). Although the northeast region is second smallest in terms of numbers of counties, it has the highest average population and a high proportion of counties in the higher urbanicity strata. In contrast, the north central region has over twice as many counties as the northeast, but only half the average population, since this region is far more rural in character. These contrasts illustrate the potential confounding of urbanicity with geographical region and the consequent value in dividing the county into regions before proceeding with the urban adjustment. Note in Table 1 that the range of most urbanicity strata is 2.5, so each successive urban stratum corresponds roughly to a 4:1 change in equivalent population size (mean squared population of the three largest places).

A model for adjustment

These regions and urban strata are used to fit a mortality rate for each stratum \(\times\) region combination (20 \(\times\) 7 = 140 parameters) in the form:

\[r_{ijk} = \lambda_{ij} \varepsilon_{ijk}, \quad i = 1, \ldots, 7 \text{ regions}, \quad j = 1, \ldots, 20 \text{ strata}, \quad k = 1, \ldots, n_{ij} \text{ counties},\]

where \(r_{ijk}\) is the age-adjusted mortality rate in county \(k\) of stratum \(j\) in region \(i\), \(\lambda_{ij}\) is the urban effect for that stratum \(\times\) region combination, \(n_{ij}\) is the number of counties in stratum \(j\) of region \(i\), and \(\varepsilon_{ijk}\) is the remaining variation that is unexplained by the urban effect. Based on other studies (e.g. ref.\(^{21}\), p. 57), a multiplicative
model in the rates, or linear in the logarithm of the rates, seems appropriate.

3.4. Model parameter estimates
For reasons explained below, the parameters $\lambda_{ij}$ are estimated by a weighted average of the county age-adjusted rates in the stratum, where the weights are the sums of the mid-year populations corresponding to years over which the rate has been averaged. These weights may be thought of as non-age-adjusted person-years for the counties. In fact, this weighted average of the rates corresponds roughly to a weighted sum of deaths in each county, divided by the total population in the stratum, where the weights equal the proportion of the standard (1970) population in each age group divided by the proportion of the county population in that age group. This analogy is outlined in greater detail in Appendix 1. If we denote by $D_{klm}$ and $N_{klm}$ the numbers of deaths and person-years, respectively, in county $k$ ($k = 1, \ldots, n_j$) in year $l$ ($l = 1970, \ldots, 1979$ or $l = 1980, \ldots, 1987$) for age interval $m$ ($m = 1, \ldots, 18$ five-year age intervals), then the estimate of the rate in an urban stratum having $n_j$ counties is defined by:

$$\hat{\lambda}_{ij} = \sum_{k=1}^{n_j} N_{klm} r_{klm} / \sum_{k=1}^{n_j} N_{klm}$$

(2)

where the ‘+’ subscript denotes summation over the relevant index.

Standard errors of the stratum averages defined by (2) can be calculated simply as

$$SE (\hat{\lambda}_{ij}) = \left[ \sum_{k=1}^{n_j} N_{klm}^2 Var(r_{klm}) / \sum_{k=1}^{n_j} N_{klm}^2 \right]^{1/2}.$$
RESULTS

Region rates

To focus on regional differences only, a population-weighted average mortality rate over all counties in a given region can be calculated. These region rates can be compared across gender and race groups as well as between regions (Figure 3). The region rate is calculated as in (2), but the index \( k \) ranges over all counties in the region, without regard to urban strata (e.g. \( k = 1, \ldots, 247 \) for the northeast).

Urban stratum rates

Figures 4–10 plot the stratum rates with limits of two standard errors for each of the seven regions for white females 1970–1979, white females 1980–1987, white males 1970–1979, and white males 1980–1987. In each plot, the stratum rate \( \hat{\lambda}_{ij} \) for stratum \( j \) of region \( i \) is plotted at \( u_{ij} \), the median urban index for the counties in that stratum, with plot character \( n_{ij} \), the number of counties in the stratum. To assess the nature of the relationship between the mortality rate and the urban index, the stratum rate \( \hat{\lambda}_{ij} \) for stratum \( j \) of region \( i \), is smoothed as a function of the urban index \( u_{ij} \). The smoothed rate, \( \hat{\lambda}_{ij} \), is found by

\[
\hat{\lambda}_{ij} = A_j + B_j (u_{ij} - 125) + \text{smooth(residual)}
\]

where

\[
\text{residual} = \hat{\lambda}_{ij} - A_j - B_j (u_{ij} - 125), \quad A_j = \text{centercept at urban index 125 for region } j, \quad B_j = \text{slope for region } j,
\]

and the ‘smooth’ of the residuals from the straight line fit (fitted by iteratively reweighted biweight regression, see ref.23 Ch. 14) is the result of applying a flexible non-linear smoother (first, 4(3RSR)2HT (ref.24 Ch. 16), and then resmoothed using Friedman’s super smoother;25 although various smoothers, both linear and non-linear and combinations thereof, were applied, this particular combination worked well in terms of both faithfulness to the underlying trend and degree of smoothness, and, moreover, both smoothers are programmed functions in S26 and S-Plus27).

This model allows for an interpretation of the values of the centercepts \( A_j \) and the slopes \( B_j \). \( A_j \) represents the approximate level of the rates across the urban index scale, and \( B_j \) represents roughly the rate of increase across this scale. The slopes can be interpreted more directly by expressing them as a percentage of the centercepts: \( (20 - B_j/A_j) \times 100\% \), the per cent change in rates for a 4:1 ratio in equivalent population size, calculated at the centercept. (Standard errors for \( A_j \) and \( B_j \) were obtained from classical weighted least squares, where the weights are those from the final iteration of the biweight regression procedures, and thus may be biased downward by a few per cent.) The standard error for the per cent change in rates \( (2000 B_j/A_j) \) is derived using standard propagation of error formulas.28 Figure 11 shows the smoothed urban trends, \( \hat{\lambda}_{ij} \), for all seven regions, for the four sets of rates.

DISCUSSION

Region rates

Figure 3 shows that the region rates for females have shown much higher per cent increases than for males, and that the increases differ by region. Also, the overall mortality rates in the region follow the same general order from one decade to the next, but the order is different for males and females. In both decades, rates for females are highest in the West Coast region, followed by the Northeast, and then the Great Lakes/Southeast regions. Rates are lowest in the Rockies...
FIGURE 4 Plot of population-weighted stratum averages versus stratum median urban index for the Northeast region. Plot character denotes number of counties in the stratum; dashes represent 95% confidence interval under the assumption of Poisson deaths, and the solid line is a smooth fit of the stratum averages as a function of stratum median urban index (see Results). (a) White females, 1970–1979 (b) White females, 1980–1987 (c) White males, 1970–1979 (d) White males, 1980–1987

FIGURE 5 Plot of population-weighted stratum averages versus stratum median urban index for the Great Lakes region. Plot character, dashes, and solid line have same interpretation as in Figure 4. (a) White females, 1970–1979 (b) White females, 1980–1987 (c) White males, 1970–1979 (d) White males, 1980–1987
FIGURE 6 Plot of population-weighted stratum averages versus stratum median urban index for the Southeast region. Plot character, dashes, and solid line have same interpretation as in Figure 4. (a) White females, 1970–1979 (b) White females, 1980–1987 (c) White males, 1970–1979 (d) White males, 1980–1987

FIGURE 7 Plot of population-weighted stratum averages versus stratum median urban index for the North Central region. Plot character, dashes, and solid line have same interpretation as in Figure 4. (a) White females, 1970–1979 (b) White females, 1980–1987 (c) White males, 1970–1979 (d) White males, 1980–1987
Urban-rural gradients in the South Central

FIGURE 8 Plot of population-weighted stratum averages versus stratum median urban index for the South Central region. Plot character, dashes, and solid line have same interpretation as in Figure 4. (a) White females, 1970-1979 (b) White females, 1980-1987 (c) White males, 1970-1979 (d) White males, 1980-1987

Urban-rural gradients in the Rockies

FIGURE 9 Plot of population-weighted stratum averages versus stratum median urban index for the Rockies region. Plot character, dashes, and solid line have same interpretation as in Figure 4. (a) White females, 1970-1979 (b) White females, 1980-1987 (c) White males, 1970-1979 (d) White males, 1980-1987
Urban–rural gradients in the West Coast

Figure 10 Plot of population-weighted stratum averages versus stratum median urban index for the West Coast region. Plot character, dashes, and solid line have same interpretation as in Figure 4. (a) White females, 1970–1979 (b) White females, 1980–1987 (c) White males, 1970–1979 (d) White males, 1980–1987

regions in both decades and second lowest in the North Central region. In all regions, the rates have increased, sometimes by a factor of almost 2. Among white males, rates are lowest in the Rockies in both decades, and North Central rates are also low. Rates in the southern regions are the highest, whereas the West Coast rates are below average. The rates for white males have increased in all regions from one decade to the next, but by smaller percentages than those for white females.

Urban trends by region

In most cases, the urban trends in Figures 4–10 are increasing, corresponding to higher rates in more urbanized counties. The increase tends to be stronger for the five easternmost regions than in the Rockies (RK). The trends in the West Coast (WC) region show apparent increases in the lowest eight strata of urbanicity, but these increases are somewhat unstable since they are based on no more than six counties in each stratum (Table 2). In all regions, the effect of the urban trend is almost always stronger for females than for males, and stronger in the earlier decade (1970–1979) than in the latter one (1980–1987) (Table 3).

The region rates in Figure 3 can be recomputed using the urban adjustment, and the urban-adjusted region rates are given in Table 4. The relative change in the two types of rates for over half of them are more than two standard errors away from 0%, but most of the changes are small (1–6%). The largest adjustments occur in the North Central region (5–13%), particularly for white females in the earlier decade where the change is 13.1% after adjustment for urbanicity.

The magnitude of the urban-rural gradient appears to be weakening over time, as the per cent changes in Table 3 are smaller than those found for white males in 1950–1969. The decrease may be due to several factors. Increased mobility and urban expansion, as well as retirement to the rural counties, may account for a higher degree of uniformity of urban exposure among the population. A more likely explanation is the decline in smoking prevalence in recent years. Haenszel et al. noted in their survey of 1958 lung cancer deaths that the ‘urban factor’ had a much greater impact on smokers than on non-smokers. Since 1965, smoking prevalence has been declining rapidly for white males but not quite so fast for white females: the change in percentage of smokers between 1965 and 1985 is over three times greater for white males (−19.5%) than for white

...
Regional urban gradients in lung cancer mortality


females (–6.0%). These two facts together may account for a weakening of the urban-rural gradient, particularly for the recent data (1980–1987) on white males.

SUMMARY
A measure of urbanization for a county, urbanicity, that involves the populations of at most the three largest places in the county, is used to assess the impact of urbanicity on lung cancer mortality rates in males and females in two decades. The urban-rural gradient is generally significant for white males and white females in both decades, with the per cent change in the rates across the scale of urbanicity ranging from 0–14% in white males and 1–22% in white females.

Even this small adjustment for urbanicity may be helpful in eliciting geographical patterns. For example, the analysis by Brillinger of birth rates in Saskatchewan shows that the simple Poisson distribution, although not adequate for the raw data, fits very well when the rates are adjusted for even one variable (weekly effect). A small systematic component should be removed to identify further variation which becomes apparent. These urban-adjusted rates will be analysed for geographical variation within each region in the forthcoming paper.

ACKNOWLEDGEMENTS
Dr K Kafadar was supported in part by National Science Foundation Grant Number DMS 95-10435 awarded to the University of Colorado at Denver.

Dr C R Goodall was supported in part by National Science Foundation award DMS 92-08656 to the Pennsylvania State University.

Research by Goodall and Tukey was prepared in part in connection with research at Princeton University, supported by contract DAAL03-91-0138 from the U.S. Army Research Office.
### Table 3 Centercepts (at urban index = 125)* and slopes of urban parameters as a function of stratum median urban index

<table>
<thead>
<tr>
<th>Region</th>
<th>Centercept</th>
<th>Standard error</th>
<th>Slope</th>
<th>Standard error</th>
<th>% change</th>
<th>Standard error</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Lung – White Female, 1970–1979</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Northeast</td>
<td>13.06</td>
<td>0.25</td>
<td>0.079</td>
<td>0.012</td>
<td>12.03</td>
<td>1.83</td>
</tr>
<tr>
<td>Great Lakes</td>
<td>12.28</td>
<td>0.13</td>
<td>0.072</td>
<td>0.006</td>
<td>11.76</td>
<td>0.96</td>
</tr>
<tr>
<td>Southeast</td>
<td>11.75</td>
<td>0.34</td>
<td>0.087</td>
<td>0.017</td>
<td>14.85</td>
<td>2.96</td>
</tr>
<tr>
<td>North Central</td>
<td>10.10</td>
<td>0.21</td>
<td>0.111</td>
<td>0.010</td>
<td>21.94</td>
<td>2.11</td>
</tr>
<tr>
<td>South Central</td>
<td>12.59</td>
<td>0.15</td>
<td>0.099</td>
<td>0.007</td>
<td>15.71</td>
<td>1.14</td>
</tr>
<tr>
<td>Rockies</td>
<td>10.82</td>
<td>0.34</td>
<td>0.030</td>
<td>0.017</td>
<td>5.55</td>
<td>3.08</td>
</tr>
<tr>
<td>Western</td>
<td>15.63</td>
<td>0.35</td>
<td>0.069</td>
<td>0.015</td>
<td>8.86</td>
<td>1.99</td>
</tr>
<tr>
<td><strong>Lung – White Female, 1980–1987</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Northeast</td>
<td>22.53</td>
<td>0.44</td>
<td>0.090</td>
<td>0.021</td>
<td>7.95</td>
<td>1.32</td>
</tr>
<tr>
<td>Great Lakes</td>
<td>21.44</td>
<td>0.23</td>
<td>0.117</td>
<td>0.011</td>
<td>10.90</td>
<td>0.72</td>
</tr>
<tr>
<td>Southeast</td>
<td>21.14</td>
<td>0.41</td>
<td>0.126</td>
<td>0.020</td>
<td>11.90</td>
<td>0.89</td>
</tr>
<tr>
<td>North Central</td>
<td>18.90</td>
<td>0.36</td>
<td>0.169</td>
<td>0.018</td>
<td>18.56</td>
<td>1.18</td>
</tr>
<tr>
<td>South Central</td>
<td>21.38</td>
<td>0.42</td>
<td>0.098</td>
<td>0.020</td>
<td>9.09</td>
<td>1.35</td>
</tr>
<tr>
<td>Rockies</td>
<td>17.43</td>
<td>0.50</td>
<td>0.033</td>
<td>0.024</td>
<td>3.77</td>
<td>2.45</td>
</tr>
<tr>
<td>Western</td>
<td>28.60</td>
<td>0.59</td>
<td>0.032</td>
<td>0.026</td>
<td>2.21</td>
<td>1.21</td>
</tr>
<tr>
<td><strong>Lung – White Male, 1970–1979</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Northeast</td>
<td>59.44</td>
<td>0.82</td>
<td>0.170</td>
<td>0.039</td>
<td>5.72</td>
<td>1.90</td>
</tr>
<tr>
<td>Great Lakes</td>
<td>58.95</td>
<td>0.46</td>
<td>0.220</td>
<td>0.021</td>
<td>7.46</td>
<td>0.99</td>
</tr>
<tr>
<td>Southeast</td>
<td>66.57</td>
<td>0.60</td>
<td>0.127</td>
<td>0.029</td>
<td>3.81</td>
<td>1.93</td>
</tr>
<tr>
<td>North Central</td>
<td>51.52</td>
<td>0.62</td>
<td>0.246</td>
<td>0.030</td>
<td>13.42</td>
<td>1.97</td>
</tr>
<tr>
<td>South Central</td>
<td>67.68</td>
<td>0.97</td>
<td>0.129</td>
<td>0.046</td>
<td>3.81</td>
<td>1.92</td>
</tr>
<tr>
<td>Rockies</td>
<td>42.53</td>
<td>1.08</td>
<td>0.099</td>
<td>0.052</td>
<td>4.65</td>
<td>2.78</td>
</tr>
<tr>
<td>Western</td>
<td>58.97</td>
<td>0.81</td>
<td>0.074</td>
<td>0.036</td>
<td>2.52</td>
<td>1.79</td>
</tr>
<tr>
<td><strong>Lung – White Male, 1980–1987</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Northeast</td>
<td>70.47</td>
<td>0.80</td>
<td>0.111</td>
<td>0.037</td>
<td>0.32</td>
<td>1.04</td>
</tr>
<tr>
<td>Great Lakes</td>
<td>73.34</td>
<td>0.89</td>
<td>0.139</td>
<td>0.042</td>
<td>3.80</td>
<td>1.15</td>
</tr>
<tr>
<td>Southeast</td>
<td>83.38</td>
<td>0.81</td>
<td>0.139</td>
<td>0.042</td>
<td>3.80</td>
<td>1.15</td>
</tr>
<tr>
<td>North Central</td>
<td>61.08</td>
<td>0.81</td>
<td>0.337</td>
<td>0.040</td>
<td>11.03</td>
<td>1.31</td>
</tr>
<tr>
<td>South Central</td>
<td>77.71</td>
<td>1.33</td>
<td>0.059</td>
<td>0.064</td>
<td>1.52</td>
<td>1.65</td>
</tr>
<tr>
<td>Rockies</td>
<td>48.57</td>
<td>0.97</td>
<td>0.039</td>
<td>0.046</td>
<td>1.62</td>
<td>1.89</td>
</tr>
<tr>
<td>Western</td>
<td>63.34</td>
<td>1.68</td>
<td>0.220</td>
<td>0.079</td>
<td>6.95</td>
<td>2.50</td>
</tr>
</tbody>
</table>

* Urban index = 125 corresponds to median for entire country (3068 counties).

Standard errors calculated from weighted linear regression, using final weights from biweight regression procedure used to calculate centercepts and slopes.

% change corresponds to a 4:1 change in equivalent population of largest urban place.

### Table 4 Region rates and urban-adjusted region rates – White Females and White Males, Lung Cancer Mortality

<table>
<thead>
<tr>
<th>Region</th>
<th>Population-weighted rate</th>
<th>Urban-adjusted rate</th>
<th>Relative change (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Northeast</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Great Lakes</td>
<td>15.81</td>
<td>25.43</td>
<td>64.83</td>
</tr>
<tr>
<td>Southeast</td>
<td>14.48</td>
<td>24.67</td>
<td>65.40</td>
</tr>
<tr>
<td>North Central</td>
<td>14.75</td>
<td>25.06</td>
<td>70.46</td>
</tr>
<tr>
<td>South Central</td>
<td>12.11</td>
<td>21.02</td>
<td>57.21</td>
</tr>
<tr>
<td>Rockies</td>
<td>15.61</td>
<td>24.57</td>
<td>71.01</td>
</tr>
<tr>
<td>West Coast</td>
<td>11.35</td>
<td>18.32</td>
<td>44.71</td>
</tr>
<tr>
<td></td>
<td>18.54</td>
<td>29.33</td>
<td>62.00</td>
</tr>
</tbody>
</table>
REFERENCES


33 For a discussion of the challenges in determining the existence of, and assessing the degree of, extra-Poisson variation in rates such as these, see Tukey JW, ‘EE3: Ecological Exploration: Character and assessment of extra-Poisson variation’. Report based on a visit to the National Cancer Institute, 17 November 1992.

(Revised version received March 1996)
Comparing age-adjusted rate of pooled deaths with the population-weighted average age-adjusted rate

Suppose there are 18 five-year age intervals, and let $D_{klm}$ and $N_{klm}$ represent the number of deaths and number of people, respectively, in the age interval $m$ in county $k$ for year $l$. Summing over years,

\[ D_{k+} = \sum_{l=1970}^{1979} D_{klm}, \quad N_{k+} = \sum_{l=1970}^{1979} N_{klm}, \]

where the ‘+’ subscript notation indicates summation over the relevant index. Let $\pi_m^*$ be the proportion of people in age group $m$ in the standard (1970) population. The direct age-standardized rate in county $k$ is:

\[ r_k = \sum_{m=1}^{18} \pi_m^* (D_{k+} / N_{k+}). \]

Thus we define the population-weighted average age-adjusted rate, $\lambda$, as

\[ \hat{\lambda} = \sum_{k=1}^{n} N_{k+} r_k / N_{++} \]

which may be expressed as

\[ \hat{\lambda} = \left[ \sum_{k=1}^{n} N_{k+} \sum_{m=1}^{18} (\pi_m^* D_{k+m} / N_{k+m}) \right] / N_{++} \]

\[ = \sum_{k=1}^{n} \sum_{m=1}^{18} (N_{k+} / N_{k++}) \pi_m^* D_{k+m}/N_{++} \]

\[ = \sum_{k=1}^{n} \sum_{m=1}^{18} (\pi_m^* / \pi_m) D_{k+m}/N_{++} \]

\[ = \sum_{k=1}^{n} \left( \text{weighted sum of deaths over all age groups} \right) \left( \text{total population} \right) \]

where the weights are the ratios of $\pi_m^*$ and $\pi_m = N_{k+m} / N_{k+}$ = proportion of county $k$'s population in age interval $m$. Note that this weighting factor has the effect of increasing or decreasing the number of deaths according to whether the 1970 US population had a larger or smaller proportion of people in the age group than there were in county $k$. This weighted sum of deaths is then divided by the total observed population in the stratum of $n$ counties.

Two other estimators of $\lambda$ might have been chosen: (a) the antilogarithm of the average of the logarithms of the age-adjusted rates, and (b) an age-adjusted pooled rate, $\sum_{m=1}^{18} \pi_m D_{++m} / N_{++m}$, if age-specific deaths and populations are available. In practice, the difference between $\hat{\lambda}$ and either one of these alternatives is very slight, except when $n$ is very small (<5) or the age distributions among the counties in the stratum are very different. We now show that both the age-adjusted pooled rate and the population-weighted average age-adjusted rate are equivalent in expectation under the assumption of Poisson deaths having a common mean, and the variance of the latter is only slightly larger than that of the former for a site such as lung cancer which involves large numbers of deaths and population counts over the time periods of interest.

Suppose $D_{klm}$ has a Poisson distribution with mean $\lambda_m N_{klm}$ where $N_{klm}$ is known and $\lambda_m$ is the age-specific death rate, averaged over all years and common to all counties in that stratum. The age-adjusted pooled rate based on pooled deaths in the stratum is

\[ \tilde{r} = \sum_{m=1}^{18} \pi_m^* D_{++m} / N_{++m} \]
where, as above, \( \pi^*_m \) is the proportion of the total 1970 US population who were in age group \( k \) and the '+' subscript notation indicates summation over the respective index. The population-weighted average of the age-adjusted rates is:

\[
\hat{\lambda} = \sum_{m=1}^{n} N_{k^+} \pi^*_m \lambda_m / N_{++}
\]

Under the assumption of Poisson distributed deaths, \( E(D_{klm}) = \lambda_m N_{klm} \), so

\[
E(\hat{\bar{r}}) = \sum_{m=1}^{n} \pi^*_m E(D_{++m}) / N_{++} = \sum_{m=1}^{n} \pi^*_m (\lambda_m N_{++m}) / N_{++}
\]

\[
= \sum_{m=1}^{n} \pi^*_m \lambda_m
\]

\[
E(\hat{\lambda}) = \sum_{k=1}^{n} N_{k^+} E(r_k) / N_{++}
\]

\[
= \sum_{k=1}^{n} N_{k^+} \left( \sum_{m=1}^{18} \pi^*_m D_{km} / N_{km} \right) / N_{++}
\]

\[
= \sum_{k=1}^{n} N_{k^+} \left( \sum_{m=1}^{18} \pi^*_m (\lambda_m N_{km}) / N_{km} \right) / N_{++}
\]

\[
= \sum_{m=1}^{n} \pi^*_m \lambda_m = E(\alpha)
\]

Thus the population-weighted average of the age-adjusted rates and the age-adjusted average rate with pooled deaths have the same expectation. The variances of \( \alpha \) and \( \lambda \) are derived as follows:

\[
\text{Var}(\hat{\bar{r}}) = \sum_{m=1}^{n} \pi^*_m \lambda_m N_{++m}^2 / N_{++m}^2 = \sum_{m=1}^{n} \pi^*_m \lambda_m / N_{++m}
\]

\[
\text{Var}(\hat{\lambda}) = (1 / N_{++m})^2 \sum_{m=1}^{n} \pi^*_m \lambda_m \sum_{k=1}^{n} N_{km}^2 / N_{km}
\]

Since

\[
\sum_{k=1}^{n} \left( (N_{k^+} - N_{km} N_{++m} / N_{++m})^2 / N_{km} \right) = \sum_{k=1}^{n} N_{km}^2 / N_{km} - N_{++m}^2 / N_{++m} > 0
\]

it follows that \( \text{Var}(\hat{\bar{r}}) < \text{Var}(\hat{\lambda}) \). Thus, if the number of recorded deaths is subject to only Poisson variation and no other sampling or non-sampling variation, then the age-adjusted rate based on pooled deaths has smaller variance than the population-weighted average of the age-adjusted rates. The inflation factor is likely to be small because of the large numbers of deaths and populations over the time periods studied here, usually less than 1% increase. For other sites with smaller numbers of deaths occurring in the less populated age groups (such as prostate cancer), the inflation factor can be much larger.

In this paper, \( N_{klm} \) is the total county population rather than the gender-specific population, and was used to adjust male and female rates separately. Almost 98% of the US counties in 1980 had a proportion of females in the range 0.47–0.53, so such gender-specific adjustments would have only minor consequences on the results.