Proceedings of the Biometrics Section of the 2000 Annual Meeting of the American Statistical Association, 2001, p. 45-52.

PREDICTION OF INCIDENT CANCER CASES IN NON-SEER COUNTIES

Linda Williams Pickle, Eric J. Feuer, Brenda K. Edwards, National Cancer Institute
Linda Williams Pickle, NCI, EPN 4103, MSC 7359, 6130 Executive Blvd., Bethesda, MD 20892

Key Words: cancer incidence, small area estimation, mixed effects models, BRFSS

1. Introduction

Maps of mortality rates have proven useful for etiologic research and setting public health policy. For example, a map of oral cancer rates among white women for 1950-69 showed a dramatic clustering of high rates in southeastern states (Mason 1975). A follow-up study in North Carolina found that this was due not to an occupational exposure, as had been suspected, but to snuff dipping (Winn 1981). Subsequently, laws were changed prohibiting the sale of smokeless tobacco to minors.

Mortality patterns result from a combination of disease occurrence (incidence), screening (if possible), diagnosis and treatment. Therefore public health researchers usually prefer to examine patterns of incidence, which are closer to the patterns of the underlying disease. Unfortunately, we are unable to produce national incidence maps of cancer because there is no nationwide cancer registry in the U.S.

The goal of a new project at the National Cancer Institute (NCI) is to predict the number of new cancer cases in the next year by state, gender, type of cancer and, in the future, race and ethnicity. In addition to their usefulness for research and policy making as noted above, predicted cancer counts are useful for:

- Cancer control – identifying where cancer screening or prevention programs are needed,
- Health resource planning,
- Cancer surveillance – identifying where or when rates are changing,
- Quality control for state cancer registries.

The primary source of information about cancer incidence in the U.S. is the NCI Surveillance, Epidemiology, and End Results (SEER) program of 11 cancer registries (http://www.seer.cancer.gov/). In 1992 Congress mandated the creation of a cancer registry in every state. The Centers for Disease Control (CDC) coordinates this effort, but to date not all states have a functioning registry and some that do are too new to have achieved a high degree of case coverage (see http://www.cdc.gov/cancer/npcr/). A list of registries certified to be of high quality is given at http://www.naaccr.org/Certification/index.html. Medicare data have been used to estimate state cancer rates, but this source obviously misses most cases under age 65 and is of limited use unless a medical procedure is specific to one type of cancer.

The American Cancer Society (ACS) annually publishes predicted numbers of new cases of the major types of cancer by gender for each state (Greenlee 2000). These are produced by multiplying the aggregated SEER registries' cancer incidence rates by age- and gender-specific nationwide populations, then deriving each state's projected incidence from its mortality count using the U.S. incidence:mortality ratio for that cancer (Wingo 1998). Counts are projected into the future using an autoregressive time trends model.

Because there is no reason to believe that incidence patterns do not vary geographically as do mortality patterns, we thought that using smaller area rates to predict the number of new cases could be more accurate than using a set of pooled rates for the entire U.S., as the ACS does. We proposed to model the relationships between cancer incidence and relevant covariate data, including mortality, in SEER registry counties, then apply this model to non-SEER counties. Predicted county counts would be summed to predict each state's number of new cases.

As a demonstration project, we planned to model data for several major cancers by gender and age for whites only. Data from a recent time period would be randomized into training and validation datasets so that model goodness-of-fit could be evaluated. If such a model could be shown to work well, rarer cancers and other racial/ethnic groups could be modeled, and time factors could be added to project the counts into the future.

Advantages of this approach to cancer prediction are:

- it utilizes the close relationship between mortality and incidence for several major cancers, and mortality data are available annually for all U.S. counties,
- prediction should be improved by including sociodemographic covariates,
- modeling at the county level allows for differences in covariate effects across geographic units.

2. Methods

The number of new cancer cases in county *i*, age group *j*, registry k, region r, denoted $d_{ij[kr]}$, is assumed to be Poisson distributed, with mean $n_{ij[kr]}\lambda_{ij[kr]}$ where $n_{ij[kr]}$ is the corresponding population at risk. (Subscripts *k* and *r* are bracketed because they are uniquely determined by county *i.)* We further assume a log-linear rate structure, i.e.,

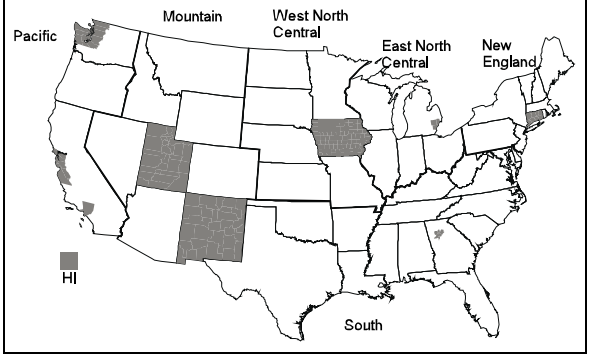
$$\ln(\lambda_{ij[kr]} \mid \beta, \gamma, \delta) = b_{0r} + f(a_j)\beta + m_{ij[kr]}\gamma + X'_{i[kr]}\delta$$

where a_j = centered age j, $m_{ij[kr]}$ is the mortality rate for county *i*, age group *j*, and $X_{i[kr]}$ is a matrix of covariates for county *i*. The regional intercepts, b_{0r}, are considered to be random effects ($b_{0r} \sim N(\beta_0, \Sigma)$). A cubic spline function of centered ages ($f(a_j)$) was necessary to accommodate downturns in some cancer rates at the oldest ages. Parameters were estimated using GLIMMIX with SAS PROC MIXED (Wolfinger 1993, SAS Institute 1999).

Regions were defined as a combination of Census Regions and Divisions so that each area contained at least one SEER registry. As seen in Figure 1, the entire South and East North Central regions are represented by a single urban area (Atlanta and Detroit, respectively). The inclusion of random effects allows information to be shared across these sparsely represented regions.

Sociodemographic variables were constructed from the Area Resource File (Bureau of Health Professions 1999) and Census data (GeoLytics Inc 1998) for urban/rural status (Butler 1994), household characteristics, income, education, occupation, medical facilities and percent population of Hispanic origin. Collinearity

Figure 1. SEER registries and regional definitions.



diagnostics were used to select representative variables from each of these broad categories to include in the model. All two-way interactions were first included in the model. Significant main effects ($p<0.05$) and very significant interactions ($p<0.01$) were retained for the final models using a backward stepwise selection process.

Because there were no areas where we know the true number of new cancer cases, two tests were planned to validate our modeling efforts. First, we randomized 50% of all individual cancer cases which occurred during 1995-96 into a training dataset and the remaining 50% into a validation dataset. Thus some cases from every SEER county were available for modeling. Secondly, we randomized the counties in each SEER registry area into two approximately equal groups and again assigned data for 1995-96 to training and validation datasets according to this county grouping. In the latter task, we were predicting counts in counties for which we had no data, similar to the ultimate project goal. These tests will be referred to as the "Case Task" and "County Task", respectively. By starting with two years of data, each training dataset had the equivalent of one year's data. Eventually we would use all 201 SEER counties to predict the other non-SEER counties, but these tests provided known results against which we could test our model. Goodness of fit was evaluated by comparing observed and predicted counts, not rates as is commonly done: $\chi^2 = \sum_{ij} \frac{[d_{ij[kr]} - E(d_{ij[kr]} \mid \beta, \gamma, \delta)]^2}{E(d_{ij[kr]} \mid \beta, \gamma, \delta) + \frac{1}{2}}$.

Cancers considered in this demonstration project were lung and bronchus, prostate, colorectal, and all other cancers for white men, and breast cancer for white women. Earlier work showed that if about 40% or more of the observations had no events (i.e., no new cases in 40% of the age-county-sex-cancer strata), the likelihood approximation might yield poor estimates (Pickle 2000). Therefore we restricted the age ranges to 25+ for breast, 35+ for lung and colorectal, and 45+ for prostate cancer; these restrictions eliminated only 0.2% of all cases.

3. Results

Not surprisingly, age was the strongest predictor of incidence for each of the cancers. Urban/rural status, education, and mortality rates were significant main effects for at least two of the five cancers examined and there were a number of significant interactions.

The observed and predicted counts were quite close for the case dataset, not only for the total number of cases among white males, but also for specific types of cancer (Table 1). A maximum of 2.5% of the

Table 1. Goodness-of-fit statistics by cancer site.

	Case Validation Dataset			%|Std. Residuals| > 2	
	# obs.	# pred.	GOF* (df)	Case Validation	County Validation
WM lung	9338	9374	738(1163)	2.4	4.6
WM prostate	17182	17305	781 (962)	2.0	1.6
WM colon	6924	6944	747(1163)	2.5	3.8
WM other	28786	28994	1351(1967)	1.7	2.1
WM total	62230	62617			
WF breast	19311	19400	991(1364)	1.6	4.0

standardized residuals were greater than 2 in absolute value. For the more difficult county task, this percentage ranged as high as 4.6%, still within the 5% expected due to chance alone.

For the five states for which the SEER program includes complete registry data, we can compare our predictions to an average of the ACS predictions for 1995 and 1996 (Wingo 1996, Parker 1997; projected from data through 1992). As seen in Figure 2a, for the total number of cases among white males, both methods were relatively accurate. However, the effect of applying an aggregate set of rates to state populations (i.e., the ACS method) is seen in Figure 2b – for a specific cancer for which there are regional differences in rates, e.g., breast cancer, a state with low rates (Iowa) is overestimated and a state with high rates (Connecticut) is underestimated.

These results suggest that the model fits are acceptable. However, plots of standardized residuals against the model covariates suggested a problem with prediction in several urban registries. In Figure 3, the

Figure 2. Observed, predicted (o), and ACS (+) estimates of the number of new cancer cases by state, 1995-96 for (a) total cancer among white males, and (b) breast cancer among white females.

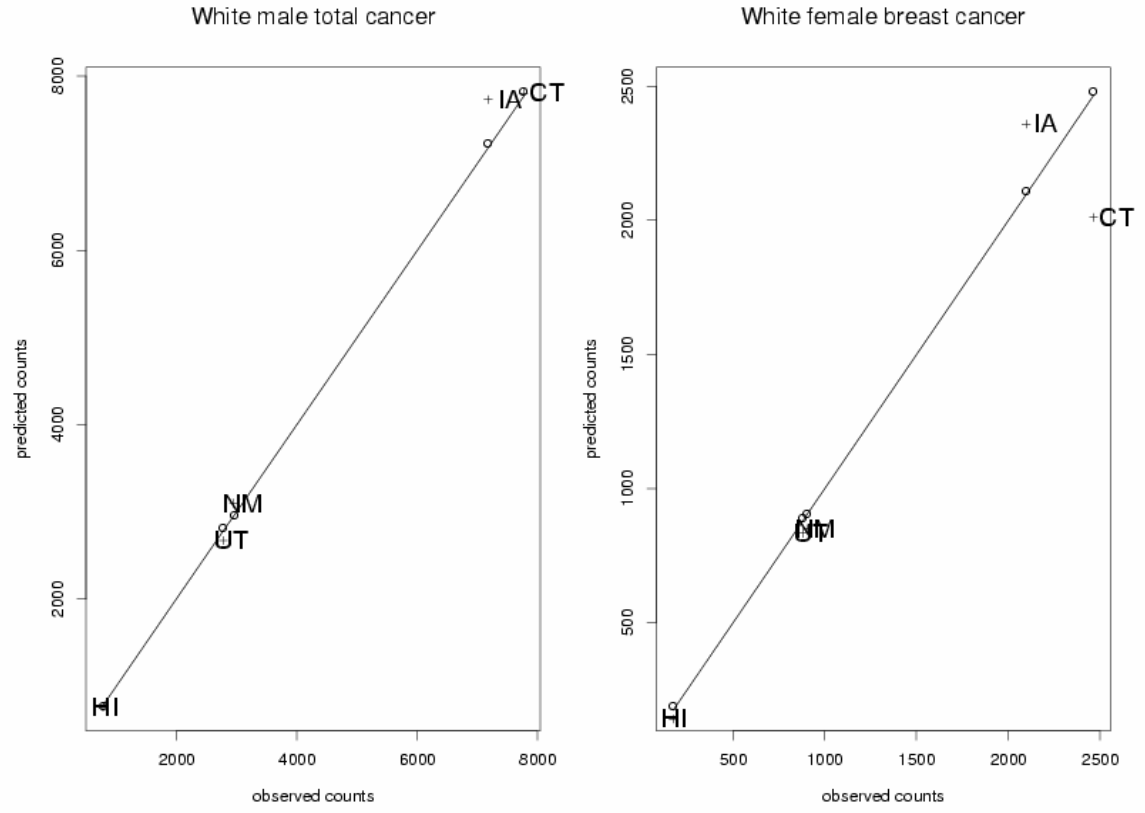
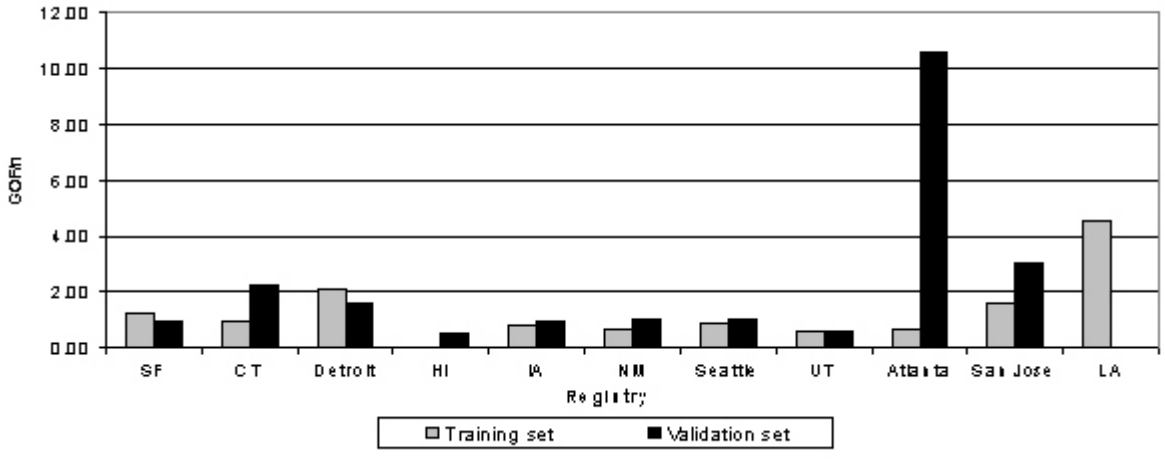


Figure 3. Scaled goodness of fit statistics by registry for white female breast cancer incidence



goodness-of-fit statistics defined above is scaled by dividing by the number of observations per registry. The expected value of this scaled statistic is between 1.0 and 2.0, depending on the number of observations. As seen in Figure 3, the Los Angeles registry was not fit well, even though its single county was randomized to the training set and thus its data was used for modeling. The prediction for Atlanta was the worst of all the registries although the training set of counties were fit well. Fulton county, where downtown Atlanta is located, was randomized to the validation dataset, while several surrounding counties were in the training set. Because Fulton county breast cancer rates are more than 25% higher than its neighboring counties, the downtown prediction was poor. Similar results were seen for the other cancers, with relatively poorer fits for Atlanta for lung cancer, Detroit for prostate, colorectal and other cancers, and Los Angeles for other cancer.

4. Conclusions from initial model

Overall, the initial models predicted the numbers of new cancer cases very well in both the training and the validation datasets. Predicted state totals differed from the ACS figures in expected ways. However, there appear to be relatively large prediction errors for several major metropolitan areas. In addition, we had hoped that mortality rates would be a stronger predictor of incidence. We were also concerned with the representativeness of Atlanta for all of the South, and so sought to improve prediction in this area.

Plans for improving these models included:

- revise the urban/rural county designation to more specifically identify large core cities,
- aggregate the mortality rates over several years for stability,
- include incidence data from 10 rural Georgia counties which are collected through the SEER program but not included in published statistics,
- add other covariates that may be more direct measures of cancer risk factors.

5. Revised model

5.1 Model modifications

The original urban/rural indicator was a four-class aggregation of the USDA codes (codes 0-1, 2-3, 4-5, 6-9) which are based on population size and metropolitan/non-metropolitan area status. For the revised model, we separated the core cities (code 0) from the other major metropolitan areas (code 1).

Mortality data were aggregated over the two available years, 1995-96, in an effort to improve the predictability of incidence. Ultimately, a number of years of mortality would be used to determine the time trends of rates.

Data from the supplemental SEER registry in rural Georgia were added (for county definitions see http://seer.cancer.gov/Registries/RuralGeorgia/). These counties have approximately a 50% white and 50% black population, with a total population of about 100,000.

A number of personal lifestyle habits have been identified as risk factors for the incidence or mortality of major diseases, including cancer. However, risk factor data are generally not available for geographic units below the state level and so cannot be applied to diseases which vary across communities. Epidemiologists may obtain information about these risk and health care utilization factors from personal interviews of individuals in very localized areas. Information about health insurance coverage or screening procedures may be available in some administrative databases, but because these sources were not set up for research purposes, the data may be incomplete.

On a broader scale, several national surveys provide estimates of risk factor prevalence, but only for large geographic regions (e.g., the National Health Interview Survey and the National Health and Nutrition Examination Survey conducted by the National Center for Health Statistics). The Current Population Survey (CPS), conducted jointly by the Census Bureau and the Bureau of Labor Statistics, includes questions about health insurance and, in a supplement conducted every few years, tobacco use. However, CPS substate estimates are only available for large population areas. The Behavioral Risk Factor Surveillance System (BRFSS) is an ongoing, telephone survey by CDC that collects and reports health risk data at the state level (http://www.cdc.gov/nccdphp/brfss/). The BRFSS data are also available at the county level, but sample sizes in any single year are too small to provide reliable estimates below the state level. We attempted to stabilize the underlying patterns in these sparse BRFSS data by aggregating the data over time and then applying a nonparametric smoothing algorithm.

5.2 BRFSS covariate development

BRFSS data for 1992-98 were aggregated by county to provide sufficient data for analysis. For each county, we calculated the mean proportions of the population, weighted by the sample design weights, who had the following risk factors:

- ever smoked 100 cigarettes,
- currently smoke cigarettes,
- obesity (body mass index (BMI) > 120% of median)
- women ages 50-64 who had had a mammogram within the last two years,
- no health plan coverage.

These questions were in the BRFSS core questionnaire, so were asked in the same way by every state. In addition, the wording of these questions was consistent throughout the period.

Fewer than 1% of the respondents did not answer the questions about smoking, mammography, or health plan coverage, whereas a BMI could not be calculated for 3.5%

of the respondents. Sixty-two of the 3098 counties had no residents sampled at all during 1992-98. These 62 missing county values were replaced by their respective weighted state means. The median number of respondents per county was 59, who represented 1514 state residents.

Because of the relatively small samples in each county, patterns in the mapped data appeared quite scattered (e.g., Figure 4). In order to reveal underlying geographic patterns in the data, we smoothed the proportions by a weighted two-dimensional smoothing algorithm (Mungiole 1999). The original value for each county was modified ("smoothed") according to the median high and low values of 30 of its nearest neighbors. The width of this smoothing window (i.e., 30 neighbors) was sufficient to begin to show regional patterns while retaining patterns that were apparent in the raw data maps. The median computation of the algorithm was weighted by county population. This ensured that unusually high or low proportions that were reliable due to large populations were not modified, whereas values based on sparse populations (e.g., those where state means were substituted) were modified to be more like those of the surrounding counties. After smoothing, clear geographic patterns emerged (e.g., Figure 5).

Although there is no "gold standard" of risk factor data with which to check our smoothed patterns, we calculated the correlations of the BRFSS variables with lung cancer mortality rates among white males, 1995-96, before and after smoothing. Correlations were stronger after smoothing than before and, of the five factors examined, current smoking was most strongly correlated with the lung cancer rates ($\rho = 0.29$). As expected, having no health plan was negatively correlated with per capita income, more so after smoothing.

5.3 Results

The modifications described in section 5.1 were implemented for white female breast cancer, including the addition of the smoothed lifestyle and health plan coverage variables to the original models. The resulting revised model included the new mammogram use variable as a significant interaction with metropolitan area. The dispersion parameter was reduced from 1.22 to 1.11, indicating that this revised model explained more of the overdispersion than the initial one. The semivariogram of the residuals shows no discernible spatial correlation remaining (Figure 6).

The scaled goodness-of-fit plot by registry (Figure 7) shows that registries that were fit well by the initial model are still fit well by the revised model. In addition, Atlanta now has an acceptable fit for breast cancer. The new rural Georgia data are fit well, but the fit for Los Angeles is not improved (data not shown).

6. Discussion

Although the initial models fit the data surprisingly well, the revised model corrects the largest error for breast cancer incidence prediction. In addition, the new lifestyle covariates are promising, e.g., mammogram use remains in the model as a significant interaction effect.

Application of the revised model to the white male data remains to be completed. Following this, a number of issues need to be addressed. For example, as we move toward publishing predicted numbers of new cancer cases, should "other" cancers be further split into specific sites? Also, cases continue to be reported to the SEER registries for several years after initial diagnosis, but we have not yet assessed or accounted for the impact of these delays on our predictions.

Our analysis so far has focused on the estimation of the number of new cancer cases. Further work remains to be done on the variance estimation for these counts. In particular, extra variation should probably be introduced for the smoothed lifestyle variables (i.e., an errors-in-covariates model).

Finally, the models need to be extended to include time trends so as to predict the future expected number of new cancer cases. For quality control purposes, counts projected one year into the future, i.e., for the next data year, are desired. However, for resource planning, for example, counts for the current or next calendar year are needed, requiring projection several years beyond the available data.

There are several potential methods for incorporating time trends into the models described here. NCI publishes some time trend statistics, such as the average annual percent change in rates and changepoints of these rates (Ries 2000, Kim 2000). Our predicted counts could be adjusted by using these independently estimated trends or a more parametric assessment of trends could be undertaken through an extension of the mixed effects models described here.

Accurate estimates of the numbers of new cancer cases by state are useful for cancer surveillance, prevention and control, for state and local resource planning, and as a quality control check for state cancer registries. We have shown that it is possible to improve upon existing methods of prediction by taking into account the geographic variation in cancer rates.

References

Bureau of Health Professions (1999), Area Resource File documentation. Washington: DHHS, Health Resources and Services Administration.

Butler MA, Beale CL (1994), Rural-urban continuum codes for metro and nonmetro counties, 1993. Washington, DC: USDA Economic Research Service Report AGES 9425.

GeoLytics, Inc. (1998), Census CD + Maps. East Brunswick, NJ.

Greenlee RT, Murray T, Bolden S, Wingo PA (2000), Cancer statistics, 2000. CA Cancer J Clin, 50(1), 7-33.

Kim HJ, Fay MP, Feuer EJ, Midthune DN (2000), Permutation tests for joinpoint regression with applications to cancer rates. Statistics in Medicine, 19, 335-351.

Mason TJ, McKay FW, Hoover R, Blot WJ, Fraumeni JF Jr. (1975), Atlas of Cancer Mortality for U.S. Counties: 1950-69. Washington: USGPO (DHEW Publ. No. (NIH) 75-780).

Mungiole M, Pickle LW, Simonson KH (1999), Application of a weighted head-banging algorithm to mortality data maps. Statistics in Medicine, 18, 3201-3209.

Parker SL, Tong T, Bolden S, Wingo PA (1997), Cancer statistics, 1996. CA Cancer- A Journal for Clinicians, 46, 5-27.

Pickle LW (2000), Exploring the spatio-temporal patterns of mortality using mixed effects models. Statistics in Medicine, 19, 2251-2263.

Ries LA, Wingo PA, Miller DS, Howe HL, Weir HK, Rosenberg HM, Vernon SW, Cronin K, Edwards BK (2000), The annual report to the nation on the status of cancer, 1973-1997, with a special section on colorectal cancer. Cancer, 88, 2398-2424.

SAS Institute, Inc. (1999), SAS/STAT User's Guide, Version 8, Cary, NC: SAS Institute Inc., pp.2083-2226.

Wingo PA, Tong T, Bolden S (1996), Cancer statistics, 1995. CA Cancer- A Journal for Clinicians, 45, 8-30.

Wingo PA, Landis S, Parker S, Bolden S, Health CW (1998), Using cancer registry and vital statistics data to estimate the number of new cancer cases and deaths in the United States for the upcoming year. J of Registry Management, 25, 43-51.

Winn, DM, Blot, WJ, Shy, CM, Pickle, LW, Toledo, A, and Fraumeni, JF, Jr. (1981), Smokeless tobacco and oral cancer among women in the southern United States. New England Journal of Medicine, 304, 745-749.

Wolfinger R, O'Connell M (1993), Generalized linear mixed models - A pseudo-likelihood approach. Journal of Statistical Computation and Simulation, 48, 233-243.

Figure 4. Observed proportion of males who currently smoke cigarettes, BRFSS, 1992-98.

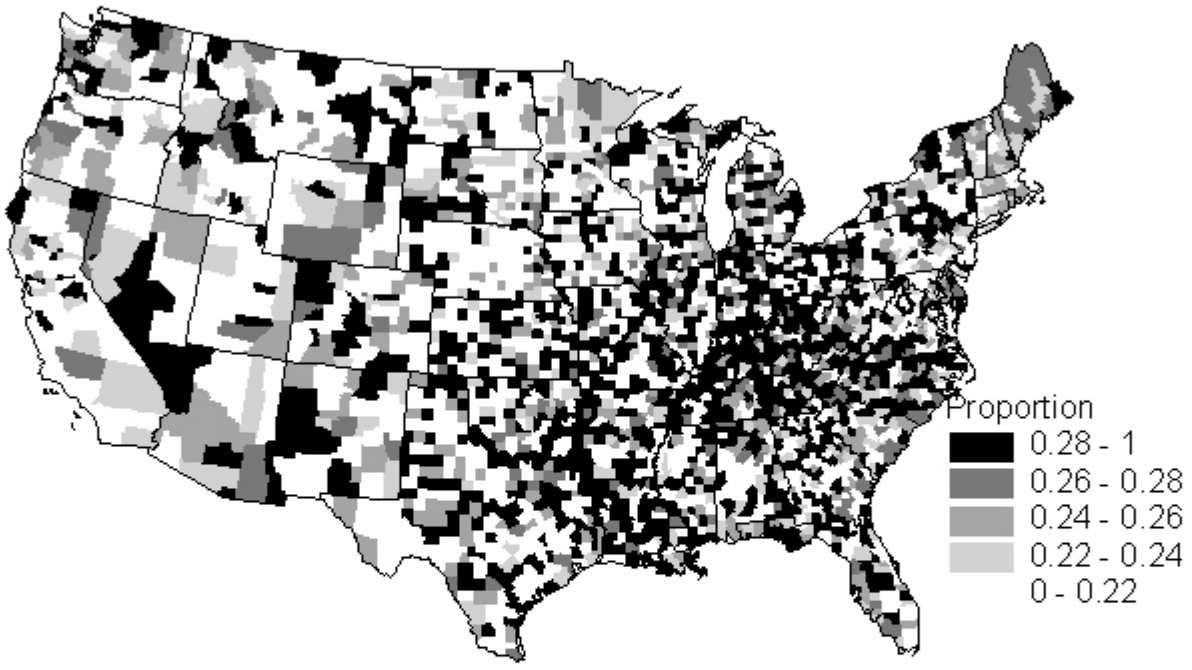


Figure 5. Smoothed proportions of males who currently smoke cigarettes, BRFSS, 1992-98.

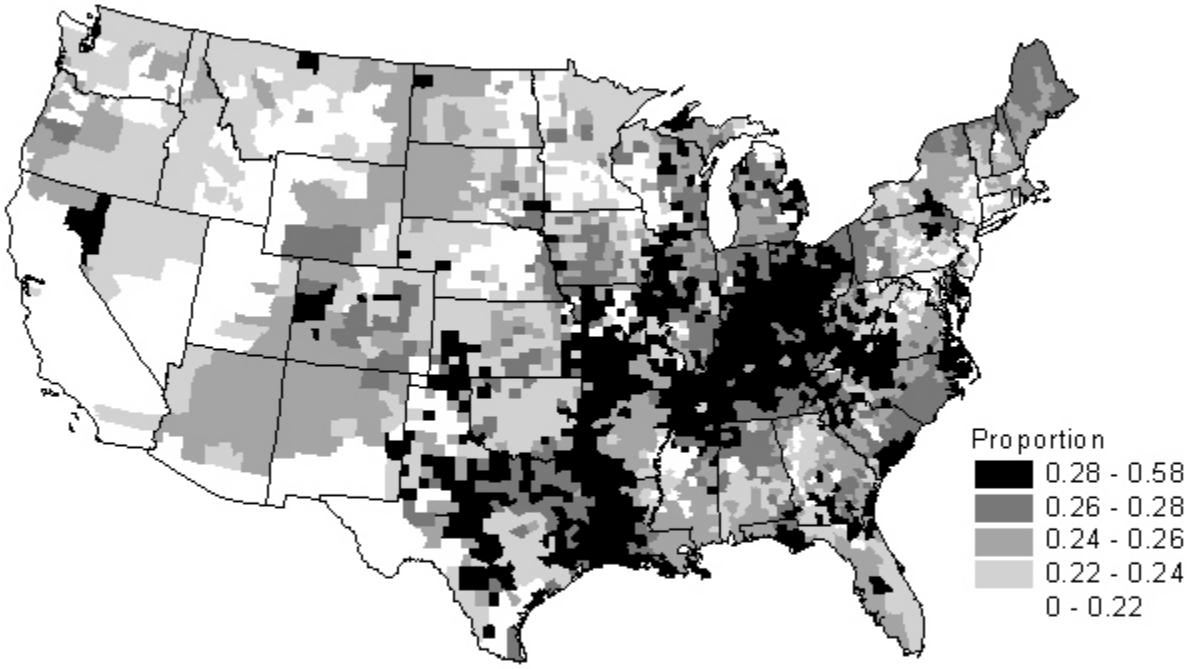


Figure 6. Robust semivariogram of residuals, breast cancer incidence among white females, 1995-96.

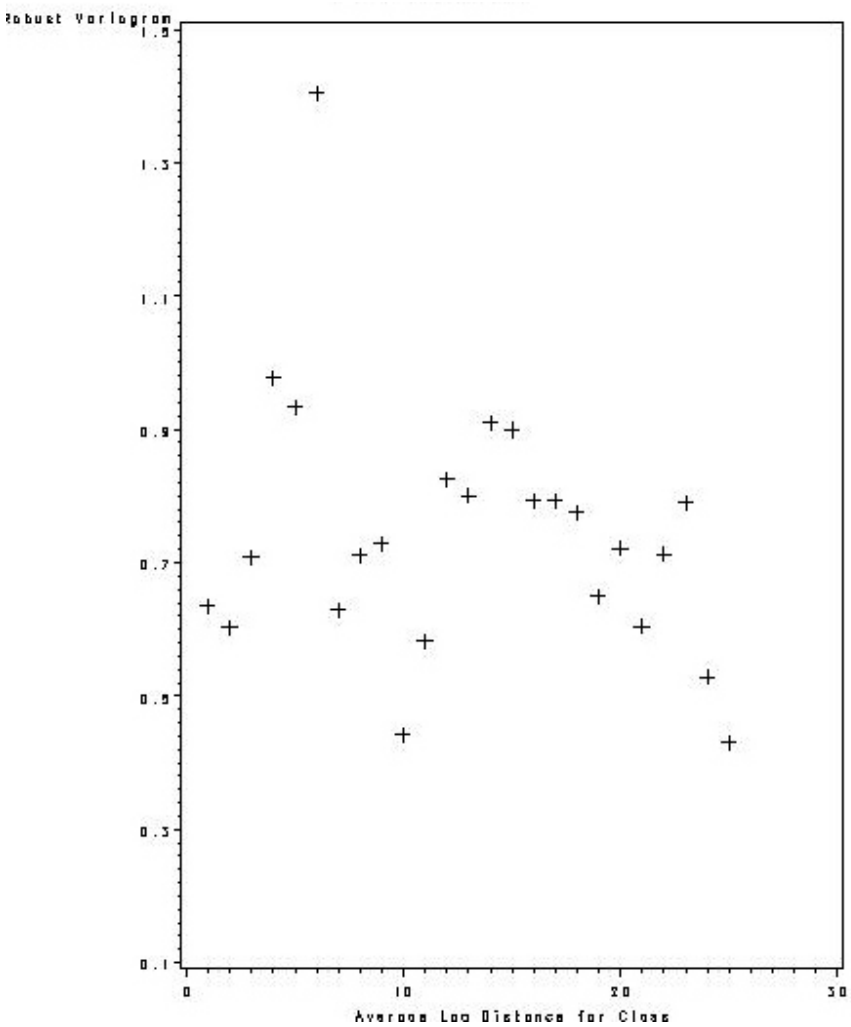


Figure 7. Scaled goodness of fit statistics by registry, validation dataset of initial (1) and revised (2) modes, white female breast cancer incidence.