BAYESIAN ANALYSIS OF CANCER MORTALITY RATES FROM DIFFERENT TYPES AND THEIR RELATIVE OCCURENCES

by

Sophie M. Delcroix

A Thesis

Submitted to the Faculty

of the

WORCESTER POLYTECHNIC INSTITUTE

In partial fulfillment of the requirements for the

Degree of Master of Science

in

Applied Statistics

February 2000

APPROVED:

Dr. Balgobin Nandram, Thesis Advisor

Dr. Hyunjoong Kim, Co-Advisor

Dr. Homer Walker, Head of Department

ABSTRACT

We analyze mortality data from prostate, colon, lung, and all other types (called other cancer) to obtain age specific and age adjusted mortality rates for white males in the U.S. A related problem is to estimate the relative occurrences of these four types of cancer.

In the recent Atlas of the Unite States Mortality (1996) each type of cancer was analyzed individually. The difficulty in doing so is that there are many small areas with zero deaths. We conjecture that simultaneous analyses might help to overcome this problem, and at the same time to estimate the relative occurrences.

We start with a Poisson model for the deaths, which produces a likelihood function that separates into two parts: a Poisson likelihood for the rates and a multinomial likelihood for the relative occurrences. These permit the use of a standard Poisson regression model on age as in Nandram, Sedransk and Pickle (1999), and the novelty is a multivariate logic model on the relative occurrences in which per capita income, the percent of people below poverty level, education (percent of people with four years of college) and two criteria pollutants, EPAPM25 and EPASO2, are used as covariates.

We fitted the models using Markov chain Monte Carlo methods. We used one of the models to present maps of occurrences and rates for the four types. An alternative model did not work well because it provides the same pattern by age and disease.

We found that while EPAPM25 has a negative effect on the occurrences, EPASO2 has a positive effect. Also, we found some interesting patterns associated with the geographical variations of mortality rates and the relative occurrences of the four cancer types.

Acknowledgments

Looking back on the time spent to accumulate the knowledge presented in this dissertation, there have been many people who contributed directly or indirectly to this work. To all of them, named or unnamed, I like to express my gratitude and deepest appreciation.

First, I would like to express special thanks to my advisor Balgobin Nandram for his endless help, guidance and patience, and for giving me the opportunity to do this project.

I also wish to thank my two co-advisors Professor Hyunjoong Kim and Linda Pickle (National Cancer Institute) for providing advice and assistance when necessary.

I would like to express my gratitude to the National Center for Health Statistics and in particularly to the Office of Research and Methodology for giving me the opportunity to work with them on this project during my summer internship.

In addition, I would like to extend a special thank to my family and friends for all their support, particularly to my sister Elise, my brother-in-law Mohamed and my special friend Maria who have been a constant source of love, reassurance, encouragement and patience.

Thank you again to everyone involved.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

Mapping of mortality rates is a valuable public health tool. The primary objectives in modeling mortality data for an atlas are to detect patterns in the mortality rates and to identify outliers from these patterns (i.e., interesting "hot-spots"). Here, we focus on cancer mortality analyses.

The 1996 Atlas (Pickle et al.) presents maps of 18 leading causes of death by sex, age, and race in the United States for the period 1988-92. This is the first publication of maps of all leading causes of death in the United States on a small-area scale. In this Atlas, information previously available only in tabular form or summarized on single map is presented on multiple maps and graphs. Broad geographic patterns by age class are highlighted by application of new smoothing algorithm, and the geographic unit for mapping is defined on the basis of patterns of health care. These new features allow public health researchers to examine the data at several different levels, to discern clusters of similar rate areas, to visualize broad geographic patterns, and to compare regional rate. With these additional tools, important geographic patterns of cause-specific mortality can be more easily identified.

The age specific numbers of deaths were modeled for each combination of race, sex, cause and place using mixed effects generalized linear models. Briefly, logarithm

of the age specific rates were modeled as a function of age, allowing each HSA to have a random slope within its particular region. Predicted age specific rates for each HSA were smoothed using a weighted head banging algorithm, with weights equal to the inverse of the rates of estimated standard errors.

Recently, there has been increased interest in inference about mortality rates for small geographical areas. Nandram et al. (1999) compared alternative models for estimating age specific and age adjusted mortality rates for all cancer for white males. They used Bayesian methods with four hierarchical models. The alternative specifications differ in their assumptions about the variation in $\log(\lambda_{ij})$ over health service areas and age classes. See also Nandram et al. (1999a) for methods used on chronic obstructive pulmonary disease. Gideon (1999) studied Bayesian methods on Poisson regression models based on the first model suggested by Nandram et al. (1999) for breast cancer mortality data. Both non spatial and spatial analyzes were investigated, by Gideon (1999).

1.1.1 Source of Data

The death counts and number at risk for this project were obtained from records of all United States death certificates in the fifty states and District of Columbia for 1988-92 and population data for 1990. The number of deaths by age, sex, place of residence, and cause of death is based on original death certificates reported to the National Center for Health Statistics (NCHS) by the States. Death certificates with age not stated were excluded, 0.025 percent of the total. Race was classified following standard procedures for United States statistics. Hispanics with no racial designation are included in the "White" category.

The population counts from the 1990 census, classified by age, race, sex, and county, were multiplied by five to create a denominator corresponding to the five years of mortality data. In few instances where the calculated number of person

years was less than the reported number of deaths, as when death occurred in a sparsely populated county before census enumeration, the years at risk were inflated to equal the total number of deaths due to any cause. The age classes are classified as 0-4 years, 5-14 years, 15-24 years, ..., 75-84 years, and 85 years and older, coded as decades 0-.25, 1, 2, 3,..., 9 (Pickle at al., 1996). Further details on the method of data collection may be found in the Technical Appendix of Vital Statistics of the United States, 1990.

The quality of the data is determined by the accuracy and completeness of the information from medical diagnosis to final coding and processing of the underlying cause of death. Beginning with mortality data for 1968, the underlying cause of death has been determined by NCHS computerized system that consistently applies the World Health Organization coding and selection rule to each death certificate using all conditions reported by certifier. Automation of these tasks and cross verification of medical conditions coding have reduced errors in assigning underlying cause of death certificate information to less than one percent. However, the completeness and accuracy of the information supplied on the certificate and the decedent's medical diagnosis remain a potential source of error.

Deaths were initially assigned to a county (or equivalent administrative unit, such as independent city or parish) according to the residence of the deceased, regardless of the place of death. There were in all 3141 geographical units, which were further aggregated into Health Service Areas (Pickle et al., 1996) by a cluster analysis of where residents aged 65 and over obtained routine short-term hospital care in 1988. An HSA may be thought of as an area that is relatively self-contained with respect to hospital care. The median number of counties per HSA is about 2 with range 1 to 20. The median number of HSAs per state is 16 with range of 1 to 58. With exception of New York City the area of each HSA is at least 250 square miles.

This project examines the geographic effects of regions as well as HSA. For this project there are twelve regions and 798 HSA, three of the nine census divisions were split to make a total of twelve regions to achieve greater homogeneity of rates (Pickle et al. 1996).

We focus our analyzes on mortality data from all cancer for white males. Cancer diseases are categorized as colon cancer, lung cancer, prostate cancer and a fourth type which we call other cancer. This fourth type includes skin, esophagus, stomach, liver, uterine cervix, multiple myeloma, lymphomas, leukemias, ovary, brain, testis, mouth, pancreas, kidney, bladder, thyroid, larynx etc.

For each type of cancer we focus on, the death rates rise steadily with age. Lung cancer has been the leading cause of cancer death in men since the 1950's and prostate cancer is primarily a disease of older men, with over 80 percent of all diagnoses occurring over age 65.

For inference on the proportional distribution of cancer types, we used the covariates income, poverty, education, and two other covariates.

The two other covariates are epapm25 and epaso2 and are called criteria pollutants. The covariate epapm25 refers to dust, dirt, smoke and other particles suspended in air. The national standard air quality includes up to 10 microns in diameter. Epapm25 includes all the particulate matters up to 2.5 microns in diameter. The particulate matters were identified as serious airbone threats to human health.

The covariate epaso2 accounts for sulphur dioxide ($S0_2$) which is closely tied to the burning of coal with a high sulphur content. It can form acid rain and has indirect health effects through contamination of surface water. Sulphur dioxide levels and particles matterns are higher in the Eastern United States.

1.2 Preliminary Analyzes of the Observed Data

In this project we focus on all cancer data categorized as colon, lung, prostate and other for white males. In this section, we perform preliminary analyzes on these data.

1.2.1 Methodology for Preliminary Analyzes

Let d_{ijk} and n_{ij} denote, respectively, the number of deaths and population at risk (number of persons years) for age class j and disease k in HSA i ($i = 1, \ldots, 798; j = 1, \ldots, 10,\ k = 1, \ldots, 4$). The age classes are $0-4$, $5-14$, $\ldots$, $75-84$, 85 and up, coded as $0.25, 1, \ldots, 9$, the midpoints of the decade intervals (*decade* $1 = .25$, *decade* $j = j - 1$, for $j = 2, \ldots, 10$).

Our model assumes that the number of deaths, d_{ijk}

$$d_{ijk}|n_{ij}, \lambda_{ij}, p_{ijk} \overset{ind}{\sim} Poisson(n_{ij}\lambda_{ij}p_{ijk}), \tag{1.1}$$

where λ_{ij} is the age specific mortality rate over all diseases, $j = 1, \ldots, 10$, $k = 1, \ldots, 4$, $\sum_{k=1}^{4} p_{ijk} = 1$, $i = 1, \ldots, 798$. Here p_{ijk} is the proportion of individuals at age j who got cancer type k in HSA i.

All the models that we consider assume (1.1).

Inference is desired for

a) the age specific mortality rate $\lambda_{ijk} = \lambda_{ij}\ p_{ijk}$ and,

b) the age adjusted rate $R_{ik} = \sum_{j=1}^{10} a_j\ \lambda_{ij}\ p_{ijk}$ where a_j are weights proportional to the U.S. population in 1940 (used in the atlas construction).

Let $\underline{d}$, $\underline{n}$, $\underline{\lambda}$ and $\underline{p}$ be the vector of the d_{ijk}, n_{ij}, λ_{ij} and p_{ijk} respectively. The joint

density of $\underline{d}$ is

$$\begin{aligned} f(\underline{d}|\underline{n},\underline{\lambda},\underline{p}) &= \prod_{i=1}^{798}\prod_{j=1}^{10}\prod_{k=1}^{4}\frac{(n_{ij}\lambda_{ij}p_{ijk})^{d_{ijk}}e^{-n_{ij}\lambda_{ij}p_{ijk}}}{d_{ijk}!}\ ,\ \sum_{k=1}^{k}p_{ijk}=1 \\ &= \prod_{i=1}^{798}\prod_{j=1}^{10}\left\{\frac{\prod_{k=1}^{4}(n_{ij}\lambda_{ij}p_{ijk})^{d_{ijk}}e^{-n_{ij}\lambda_{ij}}}{\prod_{k=1}^{4}d_{ijk}!}\right\} \qquad (1.2) \\ &= \prod_{i=1}^{798}\prod_{j=1}^{10}\left\{\frac{d_{ij.}!\prod_{k=1}^{4}p_{ijk}^{d_{ijk}}}{\prod_{k=1}^{4}d_{ijk}!}\right\}\left\{\frac{(n_{ij}\lambda_{ij})^{d_{ij\cdot}}e^{-n_{ij}\lambda_{ij}}}{d_{ij.}!}\right\} \qquad (1.3) \\ &= \left\{\prod_{i=1}^{798}\prod_{j=1}^{10}\mathrm{p}(\underline{d}_{ij}|\underline{p}_{ij},d_{ij.})\right\}\left\{\prod_{i=1}^{798}\prod_{j=1}^{10}\mathrm{p}(\underline{d}_{ij}|\lambda_{ij})\right\}. \end{aligned}$$

We note that the likelihood function can be viewed is a product of a function of the p_{ijk} and a function of the λ_{ij}. Therefore, **inference on the p_{ijk} and the λ_{ij} can be made separately**. The first is a Poisson likelihood and the second is a multinomial likelihood.

By taking the log on both parts of equation (1.2), we get

$$\begin{aligned} \log\left(f(\underline{d}|\underline{n},\underline{\lambda},\underline{p})\right) &\propto \sum_{i=1}^{798}\sum_{j=1}^{10}\left\{\sum_{k=1}^{4}d_{ijk}\ \log\left(n_{ij}\lambda_{ij}p_{ijk}\right)-n_{ij}\lambda_{ij}\right\} \\ &\propto \sum_{i=1}^{798}\sum_{j=1}^{10}\sum_{k=1}^{4}d_{ijk}\ \log\left(p_{ijk}\right)+\sum_{i=1}^{798}\sum_{j=1}^{10}\left\{(\sum_{k=1}^{4}d_{ijk})\ \log\left(n_{ij}\lambda_{ij}\right)-n_{ij}\lambda_{ij}\right\}. \end{aligned}$$

It follows that the maximum likelihood estimator of λ_{ij} is

$$\hat{\lambda}_{ij}=\frac{d_{ij.}}{n_{ij}}. \qquad (1.4)$$

and maximum likelihood estimator of p_{ijk}

$$\hat{p}_{ijk}=\frac{d_{ijk}}{d_{ij.}}. \qquad (1.5)$$

Then, the maximum likelihood estimator of the age specific mortality rate λ_{ijk} is

$$\hat{\lambda}_{ijk}\ =\ \hat{\lambda}_{ij}\ \hat{p}_{ijk}, \qquad (1.6)$$

the age adjusted rate R_i is $\hat{R}_i=\sum_{j=1}^{10}a_j\ \hat{\lambda}_{ij}\ \hat{p}_{ijk}$ where a_j are weights proportional to the U.S. population in 1940 (used in the atlas construction).

We will call $\hat{p}_{ijk}$, $\hat{\lambda}_{ij}$ and $\hat{R}_i$ the observed values of p_{ijk}, λ_{ij} and R_i respectivley.

1.2.2 Mapping the Observed Rates

Table 1.1 contains the number HSA with zero observed deaths by age-class and disease. A very large Number of HSAs contains no observed death for age classes smaller than 4. That is the data are very sparse.

Age Class	Prostate	Colon	Lung	Others
1	796	797	788	397
2	795	792	791	255
3	790	717	757	184
4	786	496	536	97
5	722	265	149	39
6	387	118	23	18
7	77	35	3	1
8	8	8	0	0
9	4	9	2	0
10	13	36	22	11

Table 1.1: Number of HSAs with Zero Observed Deaths

Age Class	Prostate	Colon	Lung	Others
1	776	469	520	47
2	722	265	149	39
3	387	118	23	18
4	77	35	3	1
5	8	8	0	0
6	4	9	2	0
7	13	36	22	11

Table 1.2: Number of HSAs with Zero Observed Deaths After Combining the First 4 Age Classes

Because of the sparseness of the data (Table 1.1), it is difficult to map age classes 1, 2, 3 and 4. We combined the first four age classes to form one so only 7 age classes remain as shown in Tables 1.2 and 1.3. This table contains also the standard million population used for age adjustment, proportional to total U.S. population in 1940.

We can see in the table that there are still many HSAs with 0 observed deaths for age class 1 and 2.

Age Class	Age (years)	Standard Population
1	0-34	594,159
2	35-44	139,237
3	45-54	117,811
4	55-64	80,294
5	65-74	48,426
6	75-84	17,303
7	85 and older	2,770

Table 1.3: Age Classes and Standard Population Used for Age Adjustment Taken from Pickle et al. (1996)

Region	1	2	3	4	5	6	7	8	9	10	11	12
# HSA	23	49	38	88	88	121	45	105	115	40	38	48

Table 1.4: Number of HSAs by Region

Table 1.4 contains the number of HSAs per region. This is a table of 798 HSAs over the continental U.S.

Since the data for the first 4 age classes are too sparse across age class for prostate and colon cancer, we defined the death rate of the first age class by the weighted average

$$\begin{cases} \lambda^*_{i1} = \frac{\sum_{j=1}^4 a_j * \lambda_{ij}}{\sum_{j=1}^4 a_j} \\ \lambda^*_{ij} = \lambda_{i(j+3)} \text{ for } j = 2, \ldots, 7 \end{cases}$$

Then, the age adjusted death rates were computed using the weights

$$\begin{cases} a*_1 = \sum_{j=1}^4 a_j \\ a*_j = a_{(j+3)} \text{ for } j = 2, \ldots, 7. \end{cases}$$

In Tables 1.5 and 1.6 we present the mean, standard deviation and median of the observed death rates over HSAs. The means for age class 1 are essentially zero and

they increase across age classes for each disease.

Age Class	All Cancer		Prostate		Colon		Lung	
	Mean	Std	Mean	Std	Mean	Std	Mean	Std
1	6.6721	3.6171	0.0041	0.0400	0.2770	0.5738	0.2046	0.4737
2	36.487	15.335	0.1403	0.7101	3.5473	4.8244	7.7909	7.2059
3	154.52	43.986	2.3117	5.3532	14.689	12.313	59.662	27.873
4	504.25	99.938	22.969	14.994	46.763	22.326	216.91	71.176
5	1091.1	169.04	108.92	35.244	108.54	37.292	438.72	119.71
6	1855.6	249.46	331.19	90.699	211.30	73.614	577.32	154.11
7	2666.5	550.65	719.81	269.48	368.29	176.94	500.28	241.34
8 (Age adjusted)	160.02	20.645	15.138	2.9028	16.076	3.8540	58.272	14.948

Table 1.5: Mean and Standard Deviation per 100,000 population for the Observed Death Rates (λ_{ijk}) over HSAs

Age Class	All Cancer	Prostate	Colon	Lung
1	6.1852	0.0000	0.0000	0.0000
2	35.250	0.0000	2.7761	7.0054
3	153.98	1.0040	13.461	57.571
4	507.02	21.910	46.879	214.22
5	1104.8	107.45	105.76	433.53
6	1869.5	328.78	209.78	580.22
7	2663.7	709.44	359.00	487.41
8 (Age adjusted)	161.82	15.034	15.860	58.037

Table 1.6: Median per 100,000 population for the Observed Death Rates (λ_{ijk}) over HSAs

Tables 1.7, 1.8, 1.9 and 1.10 present the means of the observed proportions of death p_{ijk} over HSAs by region and age class for each disease. The observed rates for prostate cancer do not vary much among the regions while they increase across age class except for region 1 and age class 1. This is due to the fact that there is little spatial clustering for prostate cancer.

Region	1	2	3	4	5	6	7
1	4.546	2.257	17.842	47.509	99.879	170.449	255.581
2	1.940	2.986	14.382	45.995	99.632	168.111	254.321
3	1.766	2.916	14.076	45.610	99.277	167.862	254.839
4	2.002	2.959	13.368	45.692	98.604	170.708	257.846
5	1.728	3.290	13.300	44.193	96.953	169.023	257.579
6	1.837	3.181	13.680	44.175	97.052	170.427	258.583
7	1.851	3.074	13.801	44.495	97.977	172.038	260.633
8	1.721	3.015	14.127	44.523	97.861	172.394	261.137
9	1.541	2.921	13.808	44.136	97.222	171.334	259.702
10	1.513	3.054	13.948	44.568	97.939	172.608	261.758
11	1.495	3.025	14.020	44.682	98.196	173.039	262.473
12	1.363	3.179	14.166	45.544	99.543	174.791	263.866

Table 1.7: Mean (10^{-3}) for the Observed p_{ijk} for Prostate Cancer by Region and Age Class

Region	1	2	3	4	5	6	7
1	38.182	82.957	91.359	102.574	122.905	134.179	151.832
2	39.525	81.723	93.953	110.361	125.177	137.503	153.308
3	40.228	83.671	95.420	107.217	120.841	133.336	149.101
4	41.035	85.538	92.407	101.107	113.425	125.835	143.655
5	42.083	88.517	90.295	97.402	109.149	123.200	142.247
6	40.402	88.743	93.055	98.829	109.744	123.815	145.022
7	40.222	89.095	93.345	99.249	109.783	123.719	144.790
8	40.038	88.328	93.767	99.185	109.541	123.685	144.661
9	39.928	88.575	92.922	97.531	106.780	121.312	142.594
10	39.392	88.950	92.687	97.735	106.465	120.621	141.924
11	39.185	88.883	92.742	97.764	106.336	120.595	141.761
12	37.754	87.482	91.651	97.192	105.828	119.485	139.642

Table 1.8: Mean (10^{-3}) for the Observed p_{ijk} for Colon Cancer by Region and Age Class

Region	1	2	3	4	5	6	7
1	39.091	203.160	343.293	385.306	342.889	288.376	179.164
2	36.615	200.691	348.815	383.950	349.294	284.484	182.140
3	37.284	203.854	356.174	393.707	358.930	290.799	183.819
4	39.748	213.306	374.410	411.430	375.226	300.825	191.101
5	39.368	219.647	383.533	423.151	387.028	307.613	192.510
6	38.828	215.159	375.936	421.382	386.156	304.683	190.719
7	38.287	214.109	373.982	419.355	384.439	302.588	188.500
8	37.613	214.741	373.785	419.727	386.249	303.508	188.482
9	36.578	214.899	374.813	422.927	390.367	307.269	190.625
10	35.926	211.417	371.709	420.396	388.985	305.784	189.657
11	35.697	210.597	371.142	419.938	388.486	305.410	189.148
12	33.559	204.696	366.529	415.858	384.794	303.653	189.752

Table 1.9: Mean (10^{-3}) for the Observed p_{ijk} for Lung Cancer by Region and Age Class

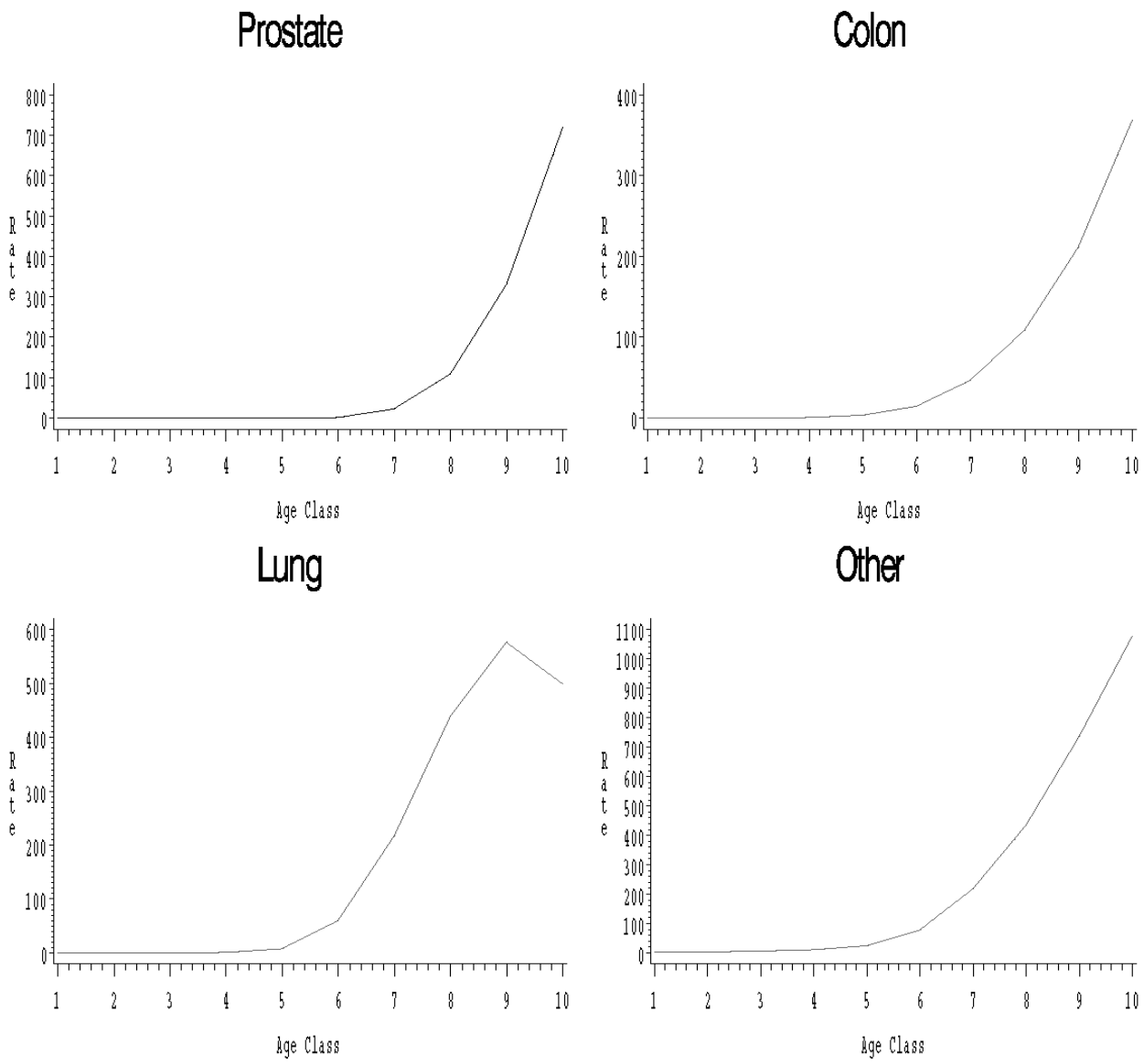
Region	1	2	3	4	5	6	7
1	918.182	711.625	547.507	464.612	434.327	406.996	413.423
2	921.921	714.600	542.851	459.694	425.897	409.902	410.231
3	920.722	709.559	534.331	453.465	420.953	408.002	412.241
4	917.215	698.198	519.815	441.772	412.746	402.632	407.399
5	916.821	688.547	512.873	435.254	406.870	400.165	407.665
6	918.933	692.917	517.329	435.614	407.048	401.076	405.676
7	919.640	693.722	518.873	436.901	407.801	401.655	406.077
8	920.629	693.916	518.322	436.565	406.349	400.412	405.719
9	921.954	693.605	518.456	435.406	405.632	400.086	407.079
10	923.169	696.578	521.656	437.302	406.611	400.987	406.660
11	923.623	697.494	522.095	437.617	406.981	400.956	406.618
12	927.324	704.643	527.655	441.406	409.835	402.071	406.741

Table 1.10: Mean (10^{-3}) for the Observed p_{ijk} for Other Cancer by Region and Age Class

Figure 1.1 presents the age specific mortality rates for prostate, colon, lung and other cancer. The death rates for prostate, colon and lung cancer are essentially zero until age class 5 where they begin to increase steadily with age class. The observed rates for lung cancer, which is the leading cause of cancer death in men, increase quickly from age class 5 to 9 where it drops. Prostate cancer being primarily a dis-

ease of older men with over 80 percent of all diagnoses occurring over age 65, is nearly zero until age class 5 an then increases steadily to be more frequent than lung cancer for age classes 9 an 10. The death rates for colon cancer increase slowly compare to prostate and lung cancers after age class 4.

Figure 1.1: Observed Death Rates for the four types of Cancer



In Figures 1.2 and 1.3 we present the maps for the observed mortality rates. The maximum likelihood estimates provide no smoothing. Apparently, there are no patterns for prostate and colon cancer for age specific rates 40 and 70. However higher adjusted rates are clustered in Mountain North and West North Central North for prostate cancer, and in the North East (age 70) and East North Central for colon cancer. For lung and all cancer, higher age specific and age adjusted mortality rates are clustered around the Appalachian region (Mississippi to West Virginia).

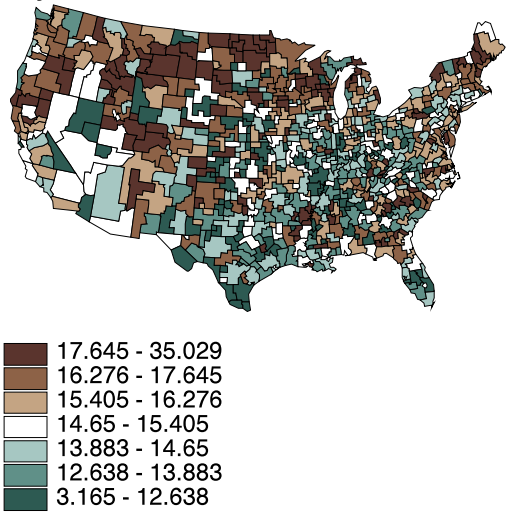
Lung cancer has been the leading cause of cancer death among men since the 1950's. Prostate cancer is primarily a disease of older men, with over 80 percent of all diagnoses occurring over age 65.

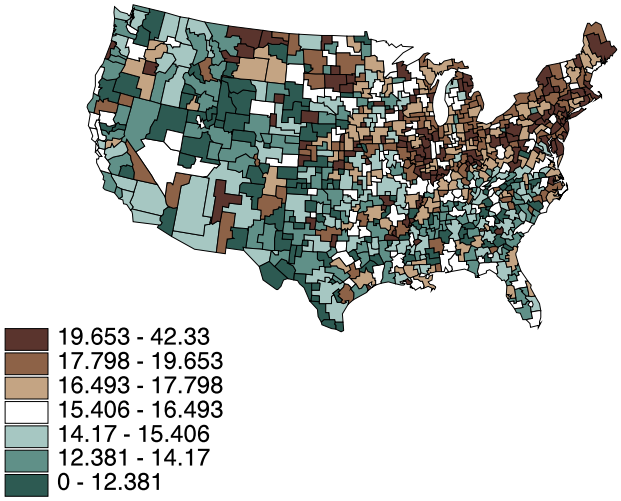
Fig. 1.2: Maps of the Observed Death Rates by Type of Cancer

Prostate

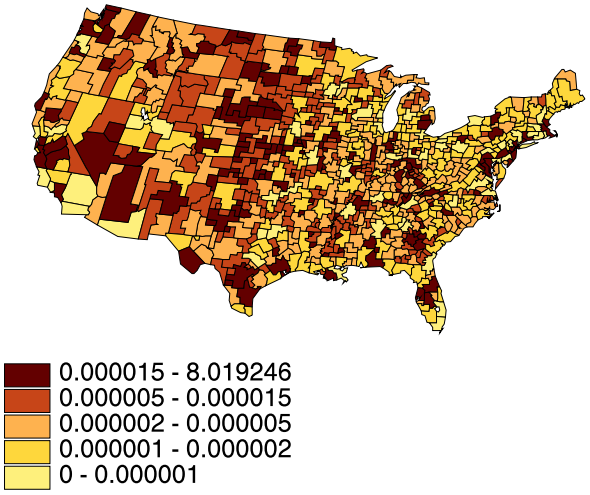
Colon

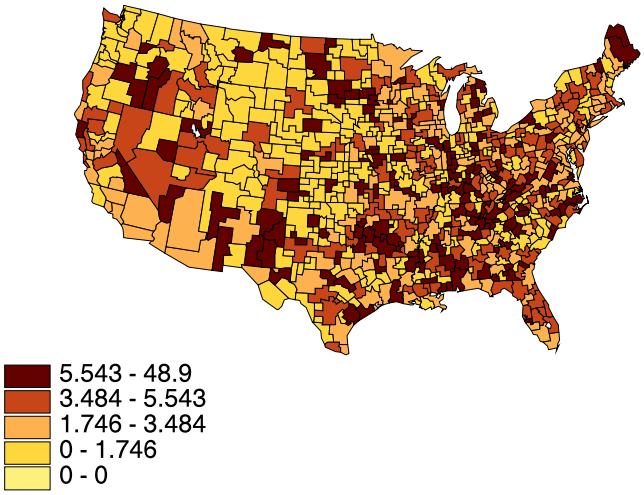
Age Adjusted



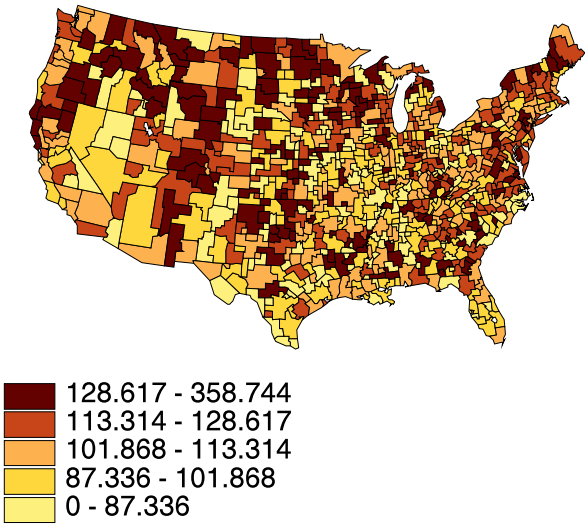


Age 40





Age 70



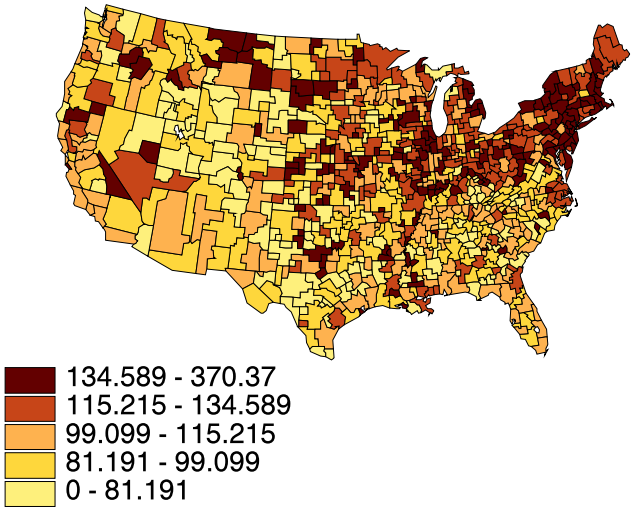
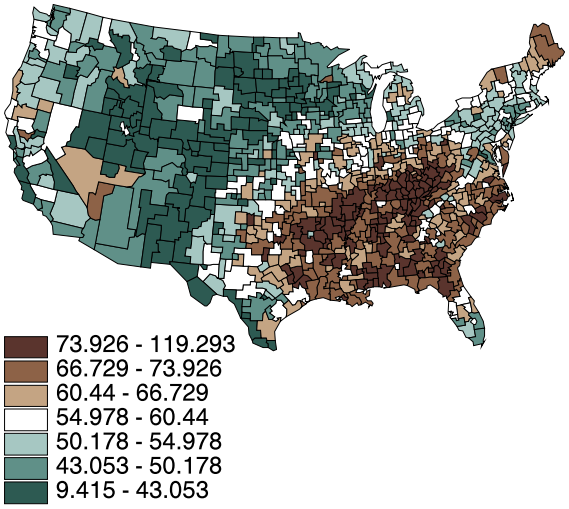
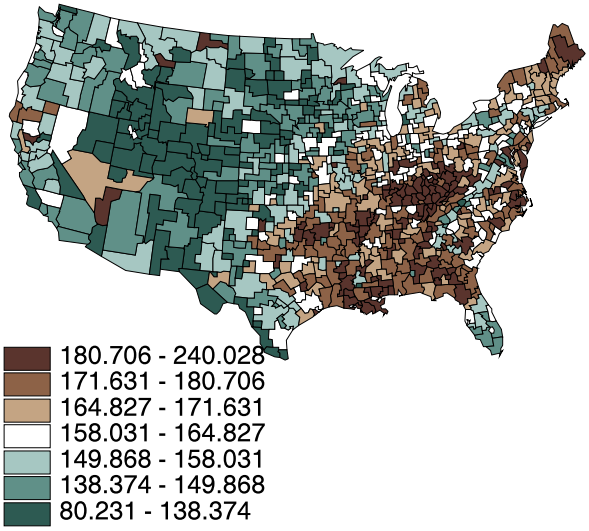


Fig. 1.3: Maps of the Observed Death Rates by Type of Cancer

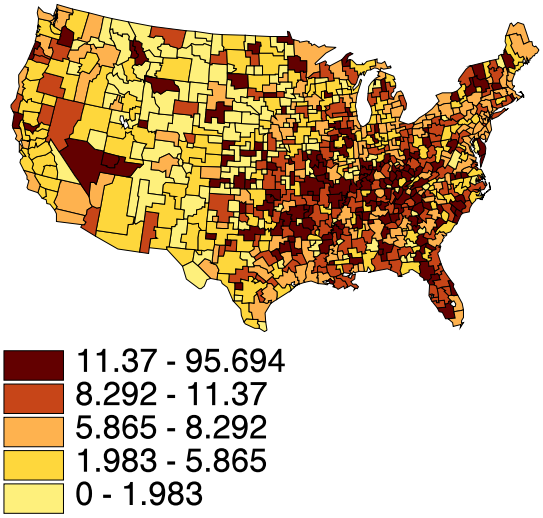
Lung | All Cancer

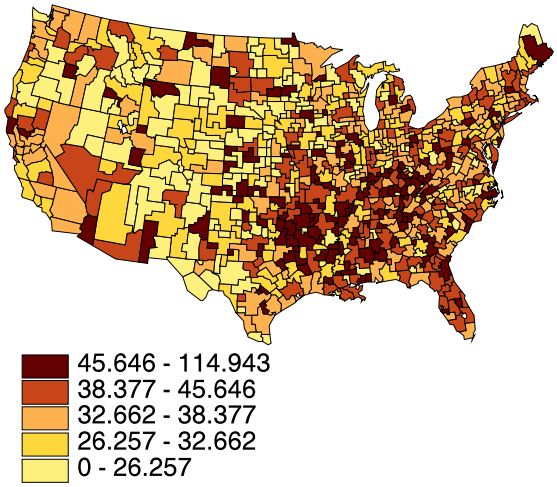
Age Adjusted



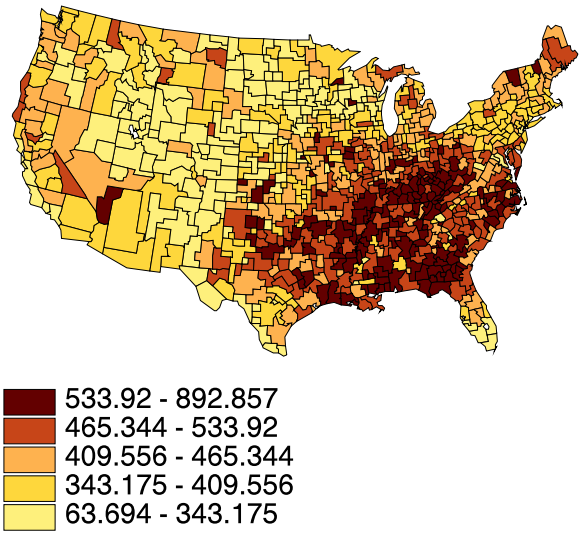


Age 40





Age 70



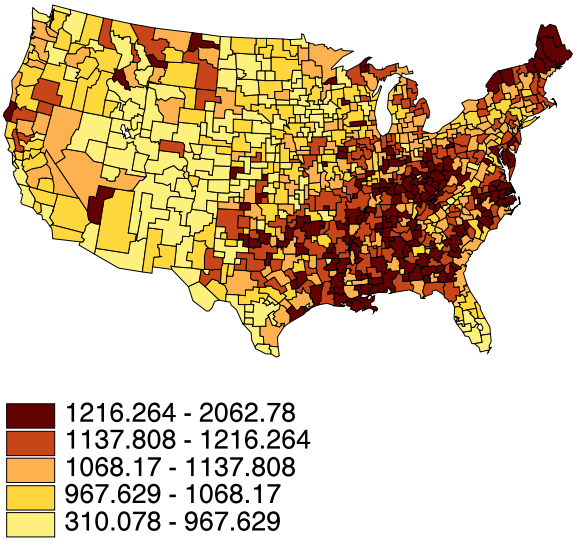


Figure 1.4 contains the box plots of the observed proportions of death by disease (p_{ijk}). For prostate cancer the proportion of death increase steadily with age classes from around 0% to 26 %. Colon cancer seems to be low for all age classes, the proportion of deaths increases slowly from 5% to 12%. The proportion of deaths for lung cancer increases steadily until age 60 from 0% to 43% and then decreases to 19% at age 85 and older. As expected, other cancer is the leading cause of death before age 30 with a proportion of about 95% and it is decreasing steadily across age class to a proportion of about 42%. Lung cancer appears to be a disease of the middle age. Other cancers affect mostly young people.

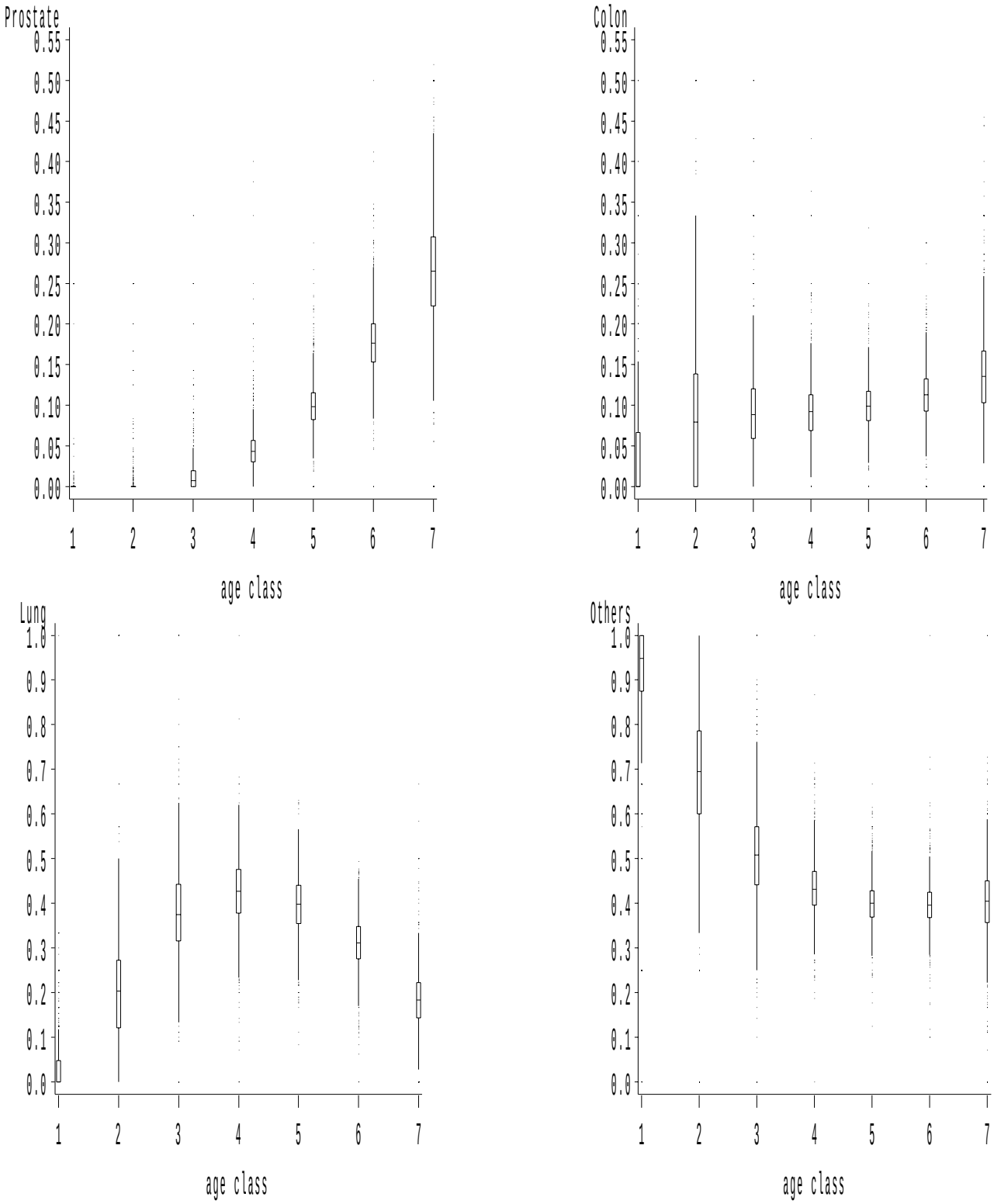


Figure 1.4: Box Plots of the Observe Proportions of Deaths by types of Cancer

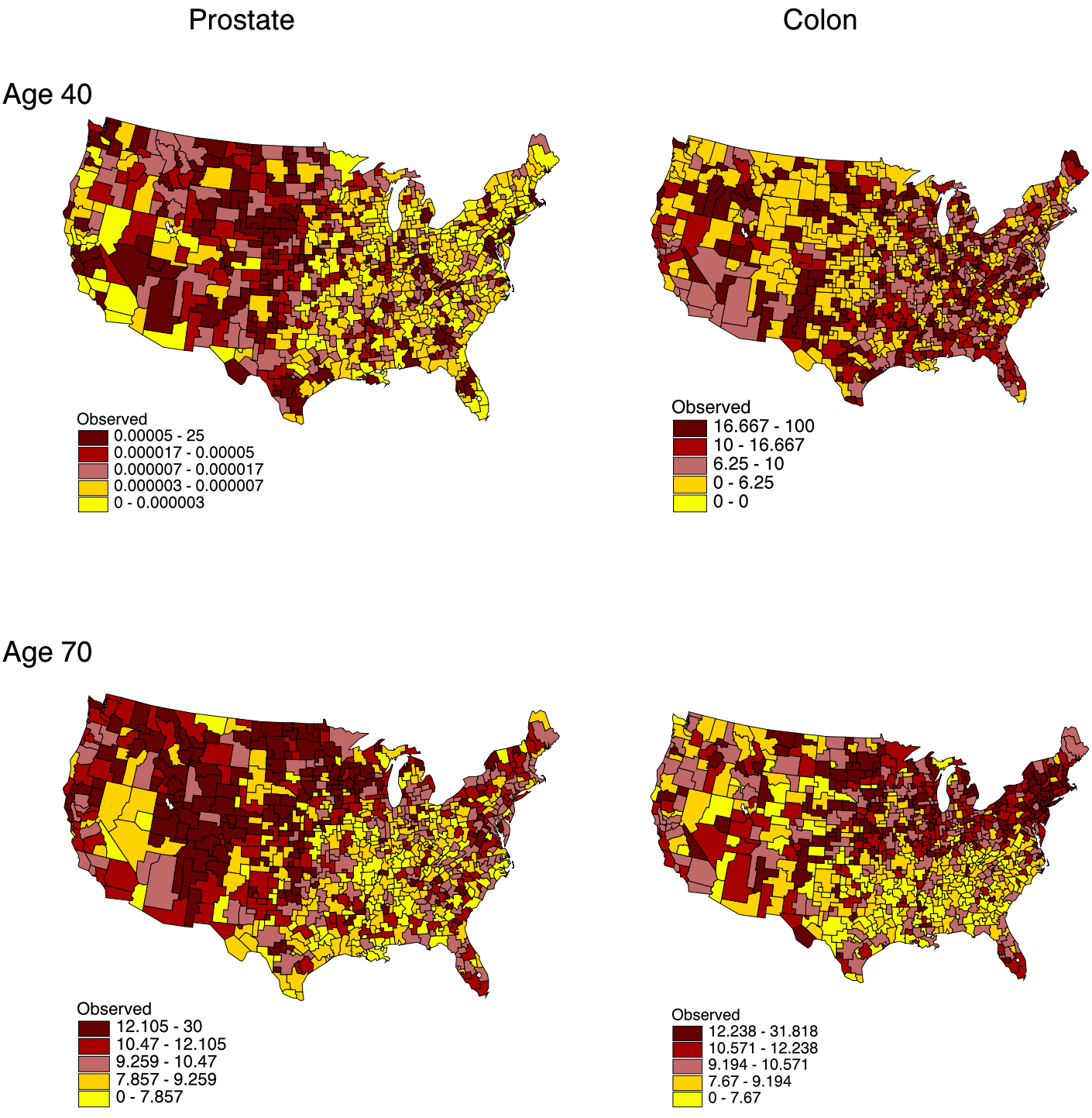
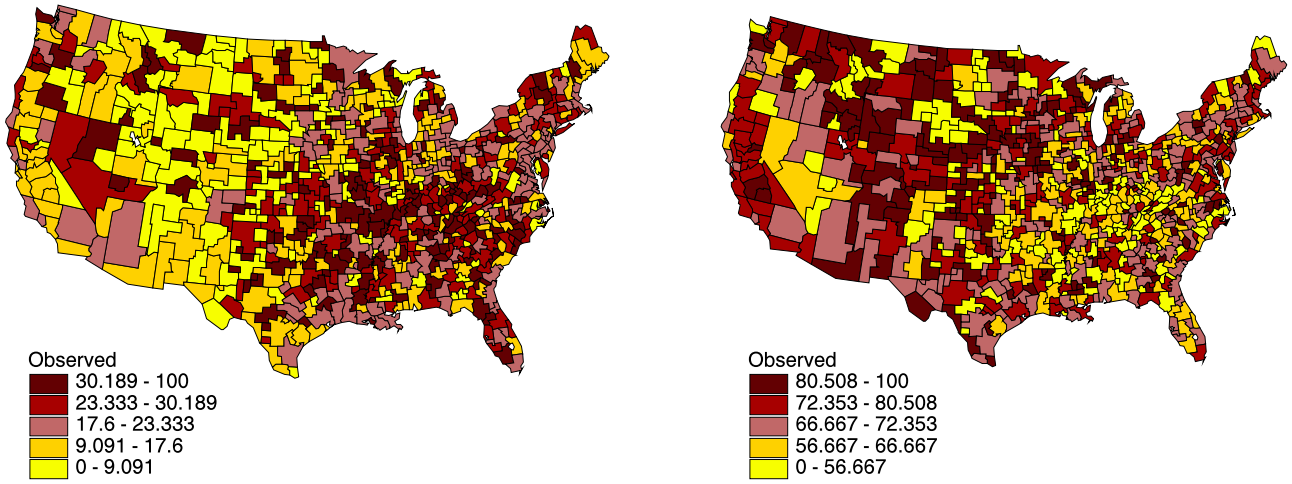


Fig 1.5: Maps of the Observed Proportions od Deaths (10-2)

Lung

Other

Age 40



Age 70

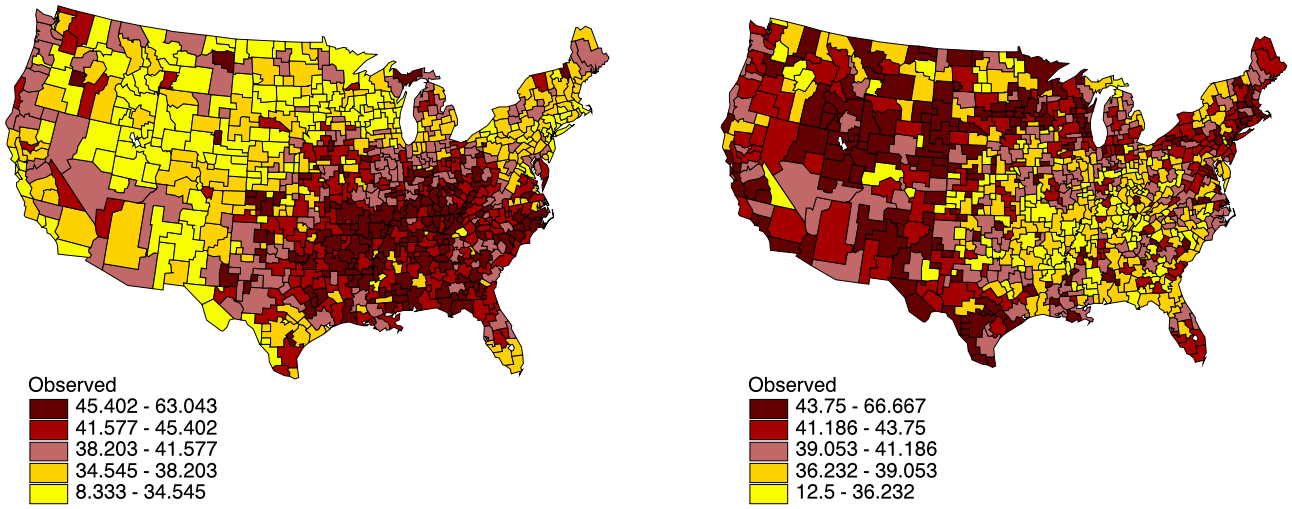


Fig 1.6: Maps of the Observed Proportions of Deaths (10-2)

In Figures 1.5 and 1.6 we present the maps for the relative occurrences. The high proportions of deaths by prostate cancer seem to be more frequent in the West side of the country while they are clustered in the East North for colon cancer. For lung cancer, higher proportions of deaths are clustered around the Appalachian region (Mississippi to West Virginia).

1.3 Model for the λ_{ij}

We use the fourth model suggested by Nandram et al. (1999) for the analysis of the mortality rates from all cancer.

We assume that $d_{ij}|n_{ij}, \lambda_{ij} \overset{ind}{\sim} Poisson(n_{ij}\lambda_{ij})$. We fit a hierarchical model with a single regression coefficient. The basis model for the analysis is as follows

$$\log \lambda_{ij} = \underline{x}_j^{'} \, \underline{\beta} + \nu_i + \delta_j \tag{1.7}$$

where $\underline{x}_j^{'} = \big(1, decade_j, (decade_j)^2, (decade_j)^3, max\{0, (decade_j - knot)^3\}\big)$ with $decade_1 = 0.25$, $decade_j = j - 1$ and for $j = 2, \ldots, 10$.

It is assumed that $\nu_i|\sigma_1^2 \overset{iid}{\sim} N\big(0, \sigma_1^2\big)$, $\delta_j|\sigma_2^2 \overset{iid}{\sim} N\big(0, \sigma_2^2\big)$ and the value of the knot that maximizes the likelihood of U.S. marginal data is 6 for "all cancer".

Here, $p(\underline{\beta}) = 1$ and $\sigma_1^{-2}, \sigma_2^{-2} \sim \Gamma(\frac{a}{2}, \frac{b}{2})$ where $a = b = 0.002$ to obtain a proper diffuse prior.

Nandram et al. (1999) showed how to fit this model for all cancer. They found that the linear structure (1.7) provided a substantially improved fit over a model in which $\sigma_2^2 = 0$. Note that $\log \lambda_{ij} = \underline{x}_j^{'} \, \underline{\beta}$ will smooth the observed data too much. Therefore, the smoothing is adjusted by adding two heterogeneous terms : ν_i and δ_j. Note only the modest ten parameters δ_j are added. We describe the Metropolis-Hastings algorithm in Appendix A.

The computations were done by region. They ran 21000 iterates, "burning in" 1000 and choosing every 20^{th} to obtain 1000 iterates which we used for output analyses. We will call this model the NSP model.

1.4 Thesis Overview

In the current chapter, we started with descriptive and exploratory data analysis, and we discussed the source of the data, data summary and pictorial representation of the data using box plots and line plots. The maps of the observed age specific and age adjusted death rates were also drawn. We observe that the data are very sparse across age classes, especially for prostate and colon cancer. A problem of weak identifiability across age class arose and was resolved by amalgamating the age classes.

In the present study we will perform Bayesian analyses of cancer mortality data by type of cancer for white males. We focus on analyzing the relative occurrences of each cancer type. Ultimately, we wish to determine patterns in the mortality data and identify outliers from these patterns (i.e., interesting "hot-spots"). We investigate Poisson regression model first on 2 approximate models and then on an exact one. We construct maps for both age specific and age adjusted mortality rates.

Since inference can be made separately on the mortality rates and the proportions of death by type of cancer, two models have to be fitted at a time. The model suggested by Nandram, Sedransk and Pickle (1999) reviewed in Section 1.3, henceforth the NSP model, is used to model the mortality rates (see Section 1.3). The objective in this project remains to investigate models for the proportions of death by disease.

In Chapter 2, we will fit two different approximate models to the data. Maps will

be constructed with the parameter estimates for age specific and age adjusted rates. We also construct box plots for the proportions of deaths by cancer types.

In Chapter 3, we apply the model first used to model the mortality rates by type of cancer. This model is not much affected by the sparseness of the data and we do not need to amalgamate the age classes. We then deduce the proportions of death by disease and fit a more elaborate model which serves as an alternative and improved model to the first approximated one presented in Chapter 2. Maps for mortality rates and proportions of death by type of cancer are obtained.

Finally, we present our conclusions, both methodological and substantive. We also present an exact model that we would have preferred to fit, but for which we experienced much difficulty with the Markov Chain Monte Carlo implementation.

Chapter 2

Approximate Models

In this chapter we describe and fit two different approximate models to the data because the data are very sparse (many HSAs have zero deaths) especially in the first 4 age classes, we amalgamated them as follows: age classes 1,2,3,4 as group 1, 5 as group 2, 6 as group 3, ..., 10 as group 7.

No new model for the λ_{ij} is discussed. We simply use the NSP model.

The two models for the p_{ijk} differ in the effects being included in this model. The second approximate model serves as an alternative and improved model to the first one. This second model does not account for age class and disease distinctively, only the interaction between age and disease remains. We also remove the intercept from the covariates for computational stability.

2.1 First Approximate Model for the p_{ijk}'s

2.1.1 Model Description

First, we assume the multinomial logit

$$\log\left(\frac{p_{ijk}}{p_{ij4}}\right) = \underline{Z}_i'\underline{\alpha} + \gamma_j + \eta_k + \delta_{jk}.$$

We use the corner point restrictions which are $\eta_3 = 0$, $\gamma_1 = 0$, $\delta_{1k} = 0$, for $k = 1, 2, 3$ (prostate, colon and lung respectively), k=4 corresponds to other cancer, $\delta_{j3} =$

0, for $j = 2, \ldots, 7$.

Then we take

$$\log\left(\frac{\hat{p}_{ijk}}{\hat{p}_{ij4}}\right) = \underline{Z}_i'\underline{\alpha} + \gamma_j + \eta_k + \delta_{jk} + e_{ijk}, \quad i = 1, \ldots, 798, \quad j = 1, \ldots, 7, \quad k = 1, 2, 3 \tag{2.1}$$

where $\underline{e}_{ij} = \begin{pmatrix} e_{ij1} \\ e_{ij2} \\ e_{ij3} \end{pmatrix} \overset{ind}{\sim} N\big(\underline{0}, \sigma^2 w_{ij}\big)_{3*3}$ and $\underline{Z}_i$ is the matrix containing the 5 predictor variables such as income, poverty, college, epapm25 and epaso2. The variable income, epapm25 and epaso2 were divided by $10,000$ for computational stability. The $\hat{p}_{ijk}$ are the MLEs with an adjustment for zeros.

In order to compute $\log\left(\frac{p_{ijk}}{p_{ij4}}\right)_{p_{ijk}=\hat{p}_{ijk}} = \log\left(\frac{d_{ijk}}{d_{ij4}}\right)$, $i = 1, 2, 3$, we used the new definition d^*_{ijk} of d_{ijk}

$$d^*_{ijk} = \begin{cases} 10^{-6} & \text{if } d_{ijk} = 0, \\ d_{ijk} & \text{otherwise.} \end{cases}$$

Moreover, for small d_{ijk}, Pickle et al. (1996) have shown that one will obtain better estimates of the λ_{ij} and the p_{ijk} using more stable quantities obtained which can be obtained computing the averages by regions for small d_{ijk}'s such as

$$d^{**}_{ijk} = \begin{cases} \frac{\sum_{i\in R}\sum_{j=1}^{10} d_{ijk}}{\sum_{i\in R}\sum_{j=1}^{10}\sum_k d_{ijk}} \frac{\sum_{i\in R}\sum_k d_{ijk}}{n_r} & \text{if } d_{ijk} < 3, \\ d_{ijk} & \text{if } d_{ijk} \geq 3, \end{cases}$$

where $R =$ region and $n_r =$ number of HSAs in the region R.

The last substitution is motivated by

$$\frac{\sum_{i\in R}\sum_{j=1}^{10} d_{ijk}}{\sum_{i\in R}\sum_{j=1}^{10}\sum_k d_{ijk}} \approx \frac{d_{ijk}}{\sum_k d_{ijk}} \quad \text{and} \quad \frac{\sum_{i\in R}\sum_k d_{ijk}}{n_r} \approx \sum_k d_{ijk} \ .$$

First we obtain the least square estimates of $\underline{\alpha}, \underline{\gamma}, \underline{\eta}, \underline{\delta}$. Let $\underline{\beta'} = \big(\underline{\alpha'}, \underline{\gamma'}, \underline{\eta'}, \underline{\delta'}\big)$, then we have

$$\hat{\underline{\beta}} = \big(X'W^{-1}X\big)^{-1}X'W^{-1}Y,$$

$$\widehat{Cov}\big(\hat{\underline{\beta}}\big) = \big(X'W^{-1}X\big)^{-1}\hat{\sigma}^2,$$

$$\hat{\sigma}^2 = \big(\underline{Y} - X\hat{\underline{\beta}}\big)'\big(\underline{Y} - X\hat{\underline{\beta}}\big)\Big/(21 \times n - 26).$$

The following identity simplifies the computations

$$X'W^{-1}X = \sum_{i=1}^{n} X_i' W_i^{-1} X_i \text{ and } X'W^{-1}Y = \sum_{i=1}^{n} X_i' W_i^{-1} Y_i \ .$$

Then we approximate the distribution of these parameters by a normal distribution

$$\hat{\underline{\beta}} \ \sim \ N_{26}\left(\underline{\beta}, \widehat{Cov\big(\hat{\underline{\beta}}\big)}\right).$$

Now pretending as though $\underline{\beta}$ has a uniform non informative prior, we have

$$\underline{\beta}|data \ \sim \ N_{26}\left(\underline{\beta}, \widehat{Cov\big(\hat{\underline{\beta}}\big)}\right). \tag{2.2}$$

It is straight forward to obtain a sample of 1000 random deviates from (2.2). By substituting these into (2.1) we obtain an approximation to the distribution of $\{p_{ijk}\}$, see Appendix A. However, this is a rough way to smooth the MLEs of the p_{ijk}.

2.1.2 Estimates and Maps

In order to draw the maps, 1000 of p_{ijk} and λ_{ij} were drawn from the starts obtained from the previous models as explained in the previous sections. One sample of 1000 death rates $\lambda_{ijk} = \lambda_{ij}\ p_{ijk}$ and adjusted death rates $R_{ik} = \sum_{j=1}^{10} a_j\ \lambda_{ij}\ p_{ijk}$ were deduced. We summarize the results by maps using only the means and standard deviations of the sample. Each age class and disease were considered separately for mapping across the U.S.

In Table 2.1 we present $E(.|\underline{d})$ and $SD(.|\underline{d})$ for the age specific and age adjusted (age

	All Cancer		Prostate		Colon		Lung	
Age Class	Mean	Std	Mean	Std	Mean	Std	Mean	Std
1	6.083	1.162	.0000	.0000	.5561	.1088	.2023	.0401
2	37.45	5.104	.0002	.0000	5.848	.8455	8.325	1.222
3	152.9	20.65	3.941	.5549	21.93	3.093	55.88	7.981
4	512.5	62.14	28.74	3.600	120.6	14.90	183.6	23.34
5	1070	149.7	115.8	16.66	241.2	34.34	358.5	52.37
6	1887	188.1	324.5	32.82	430.4	43.24	507.1	52.15
7	2724	263.6	657.3	63.80	650.6	62.94	480.1	47.33
8	160.0	17.46	15.82	1.656	34.34	3.806	50.06	5.916

Table 2.1: Mean and Standard Deviation per 100,000 population of the Death Rates (λ_{ijk}) over HSAs

class 8) rates averaged over HSAs. They look similar to the observed rates except for colon cancer for which the estimated rates are almost two times larger than the observed ones after age class 2, and for prostate cancer for which the rates are overestimated until age class 5. Based on the model, the estimated age mortality rates for age class 1 are still approximately zero. We observe that the values of the estimates increase across age class.

Table 2.2 presents the means over the 1000 simulations of the parameter estimates. The estimates follow the patterns observed in the mortality data. The estimates accounting for prostate and colon cancer are mostly negative except for prostate cancer at age class 7 (85 and older) since lung cancer is the leading cause of cancer death and the disease for which the fixed effect on disease $\eta_3 = 0$. The parameter estimates accounting for age class and disease are significant. The proportion of deaths by prostate cancer increases. For colon cancer, they increase until age class 4 (age 60) and then drop. For lung cancer, they increase with age class and drop after age class 7 (85 and older) as in the observed data. Concerning the covariates, we observe that the most significant are the intercept, epapm25 and income. The 4 remaining covariates do not seem to explain a lot of the variation. We also can notice

Covariates	Mean	Std	Ratio
Intercept	-3.049	0.2361	-12.9
% Income	-.1377	0.0291	-4.73
% Poverty	-.0082	0.0042	-1.95
% College	-.0027	0.0031	-0.87
Epapm25	-6.623	1.4569	-4.55
Epaso2	0.9549	0.3745	2.55
γ_2	2.2881	0.2363	9.68
γ_3	3.0728	0.2308	13.31
γ_4	3.3400	0.2284	14.58
γ_5	3.3273	0.2289	14.53
γ_6	3.1083	0.2295	13.53
γ_7	2.6501	0.2300	11.52
δ_{21}	-1.595	1.5512	-1.03
δ_{31}	6.9297	1.2255	5.65
δ_{41}	8.1146	1.2167	6.67
δ_{51}	8.9451	1.2141	7.37
δ_{61}	9.7457	1.2119	8.04
δ_{71}	10.639	1.2114	8.78
δ_{22}	-1.139	1.0393	-1.10
δ_{32}	-1.700	1.0369	-1.64
δ_{42}	-1.786	1.8730	-0.95
δ_{52}	-1.588	1.5987	-0.99
δ_{62}	-1.234	1.5911	-0.77
δ_{72}	-.6002	1.5928	-0.38
η_1	-10.27	1.9714	-5.21
η_2	0.3418	1.5950	.21

Table 2.2: Mean and Standard Deviation of β over the 1000 Simulations

that the parameters corresponding to the age classes are all significant. However the interaction of age class and disease is only significant for prostate cancer and after age class 1. The parameter accounting for disease is very significant for prostate but is not for colon cancer.

In Figure 2.1 we present the distributions of the proportions of deaths by disease. The box plots are similar to the one obtained from the observed proportions in Figure 1.4 for prostate, lung and other cancers. However the pattern for colon cancer is quite similar across age class but the proportions are overestimated.

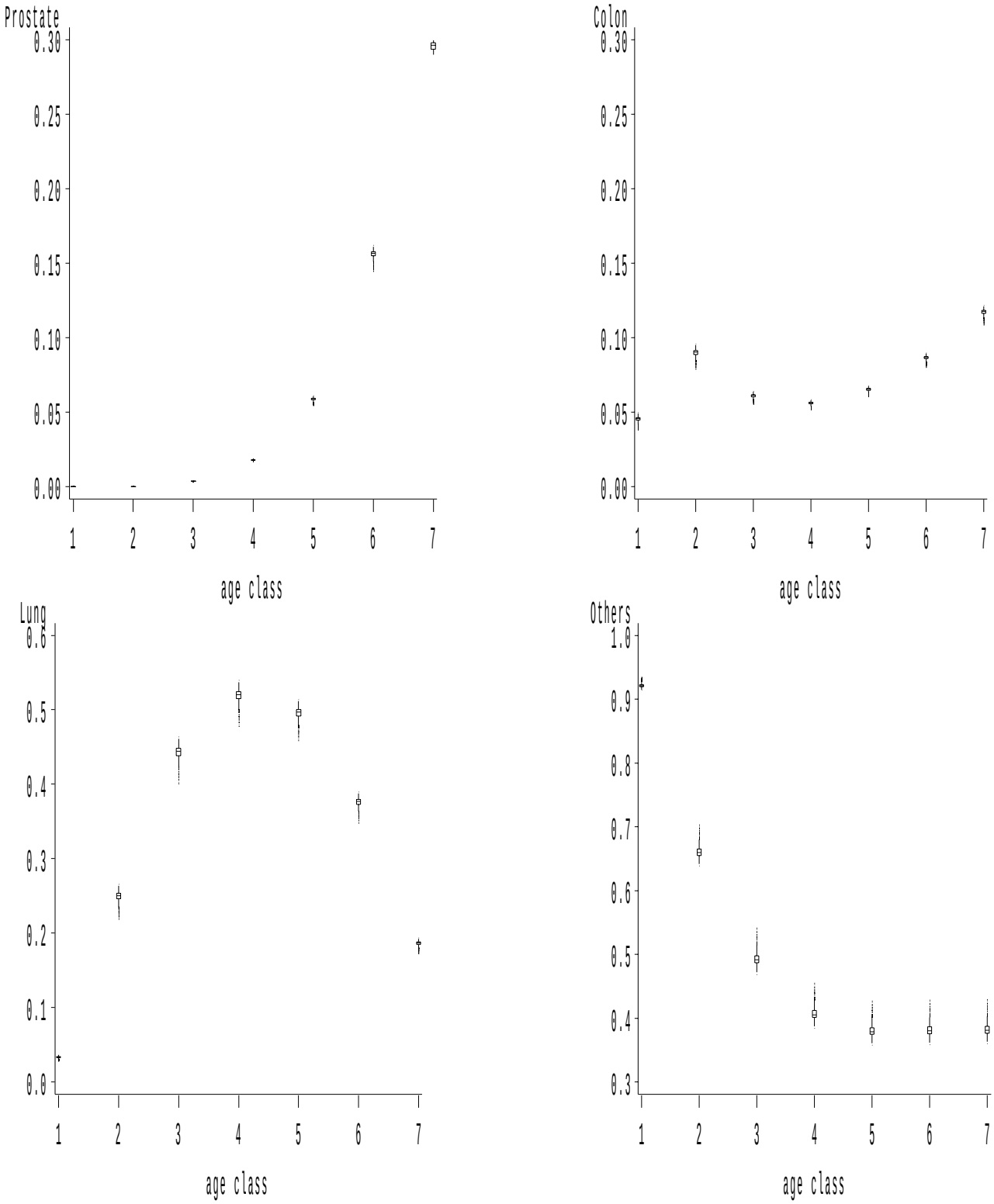


Figure 2.1: Box Plots of the Proportions of Deaths by Disease

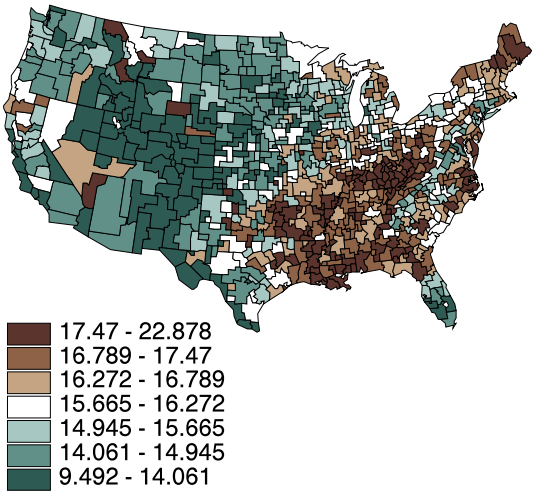
In Figure 2.2, 2.3, 2.4 and 2.5 we present the maps of the mortality rates (for age specific 40, 70 and age adjusted) and of the relative occurrences respectively. The same pattern is observed for each disease and each age class as a concentration of high mortality rates around the Appalachian region (Mississippi to West Virginia). We observe that the mortality rates obtained for colon cancer are slightly different from the observed ones as in Table 2.1.

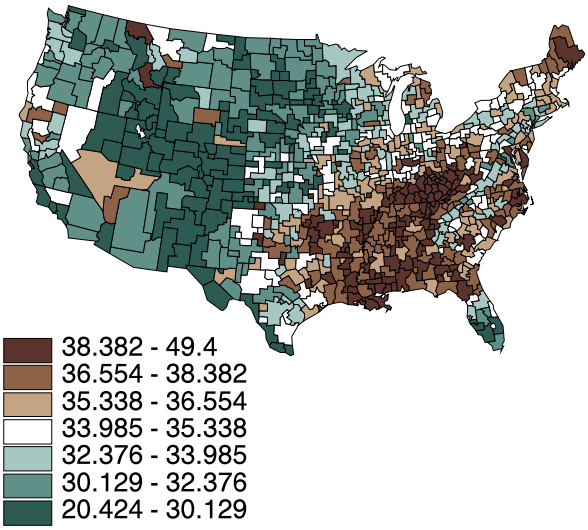
Fig. 2.2: Maps of the Estimated Death Rates by Type of Cancer

Prostate

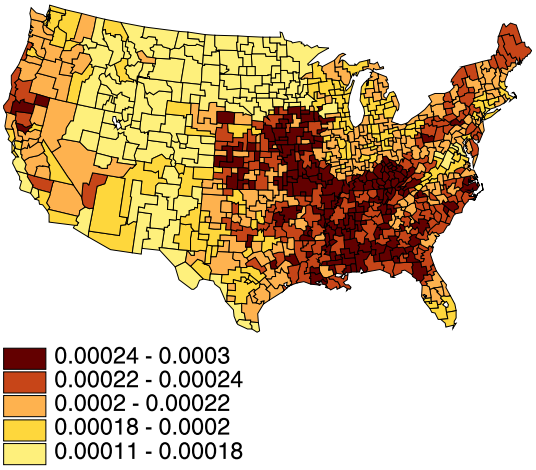
Colon

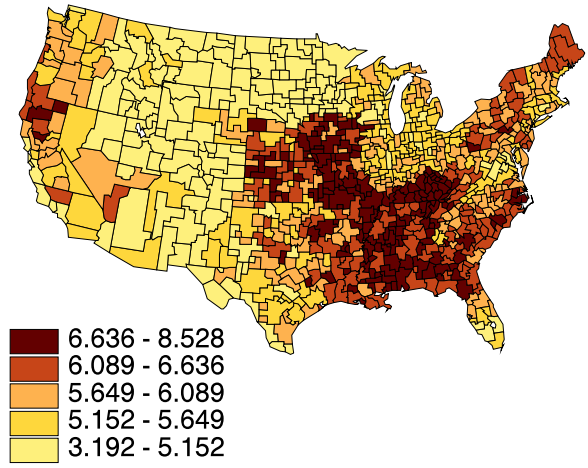
Age Adjusted



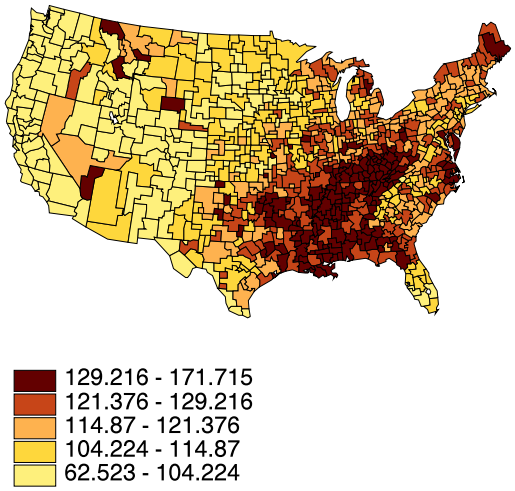


Age 40





Age 70



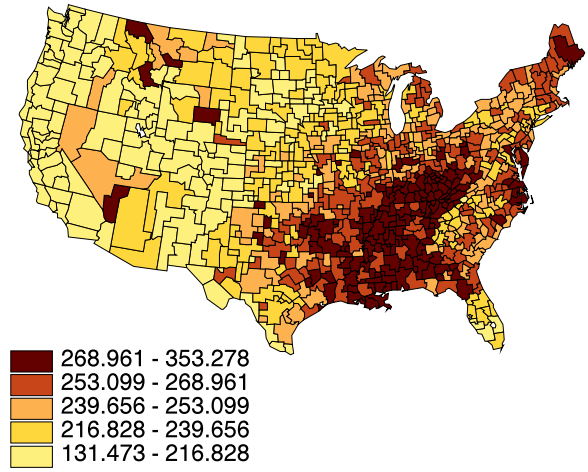
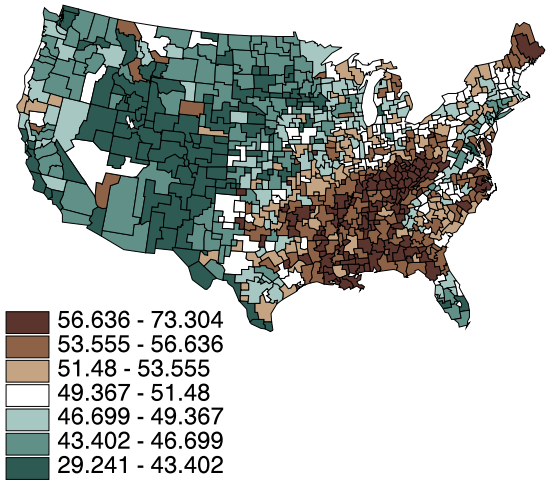


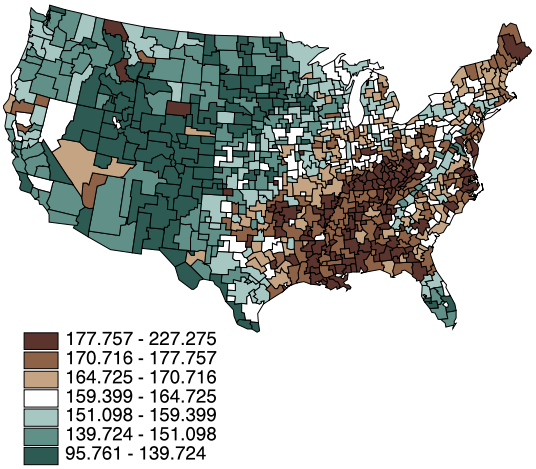
Fig. 2.3: Maps of the Estimated Death Rates by Type of Cancer

Lung

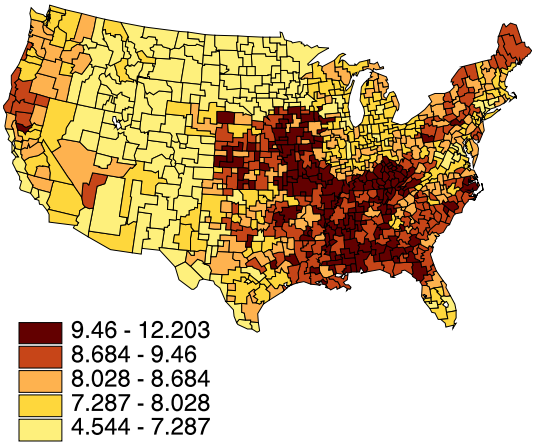
All Cancer

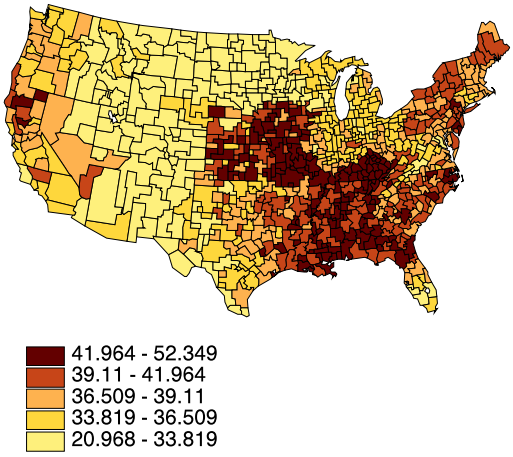
Age Adjusted



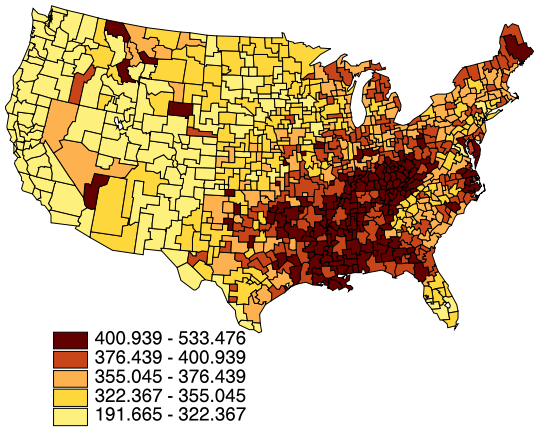


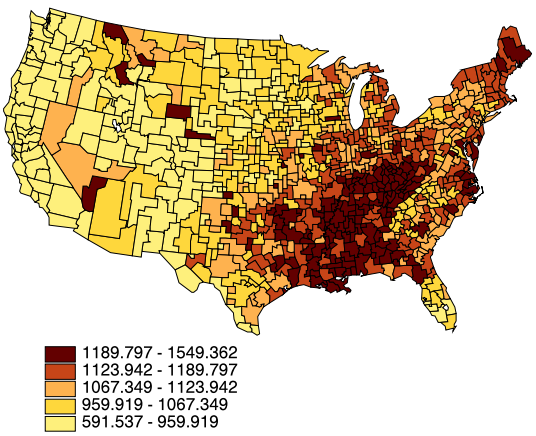
Age 40





Age 70

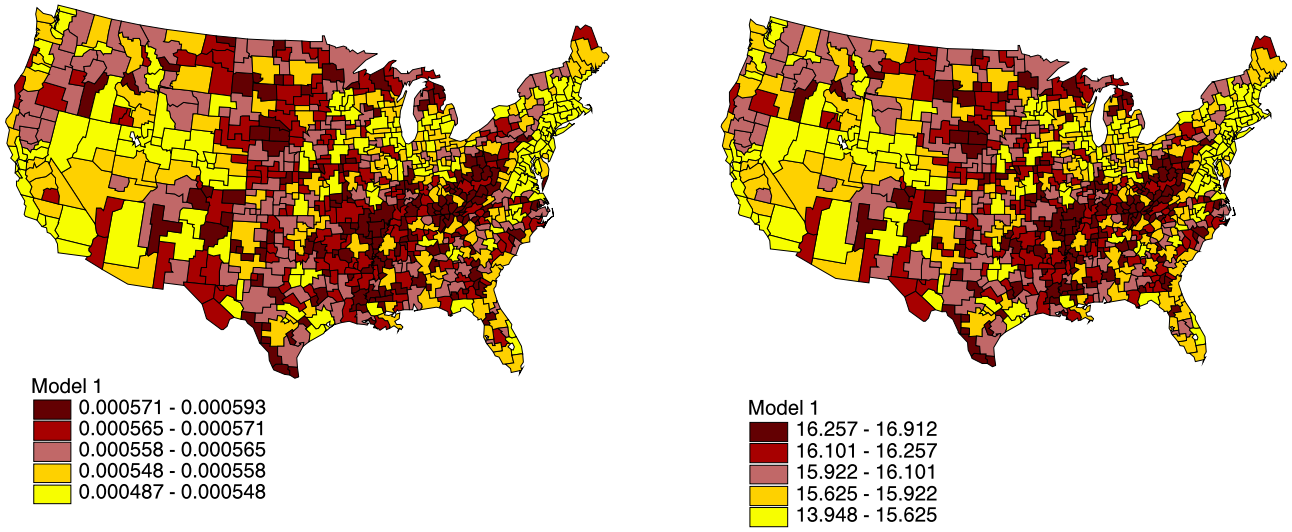




Prostate

Colon

Age 40



Age 70

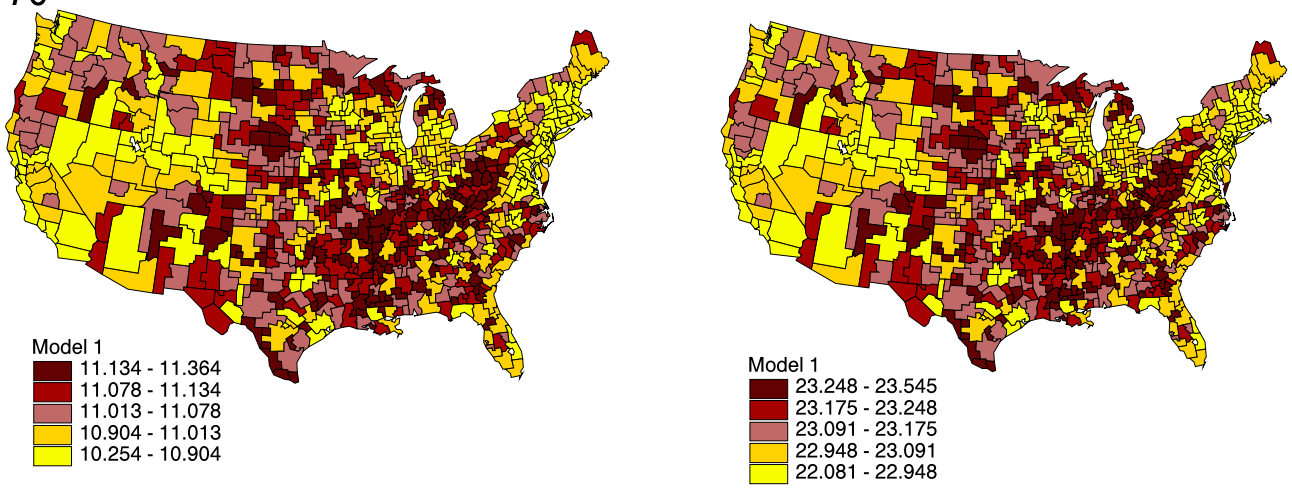
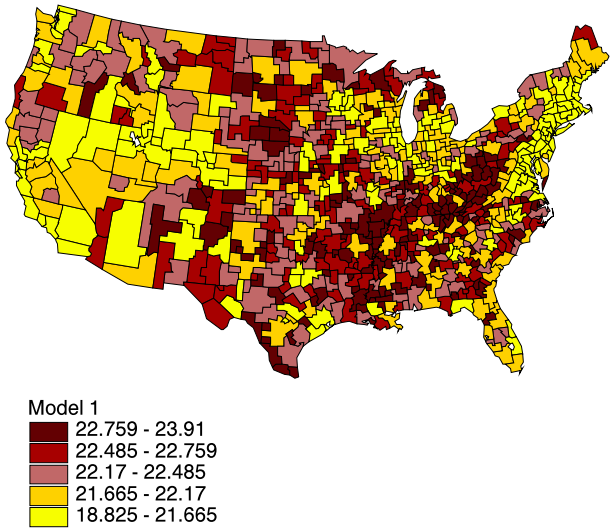


Fig 2.4: Maps of the Proportions of Deaths (10-2)

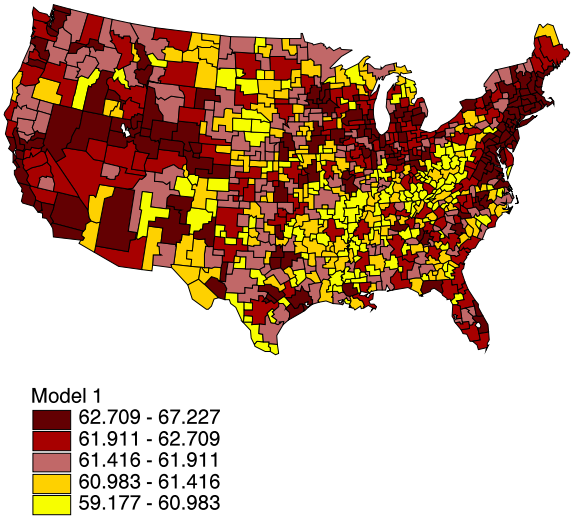
Lung

Other

Age 40

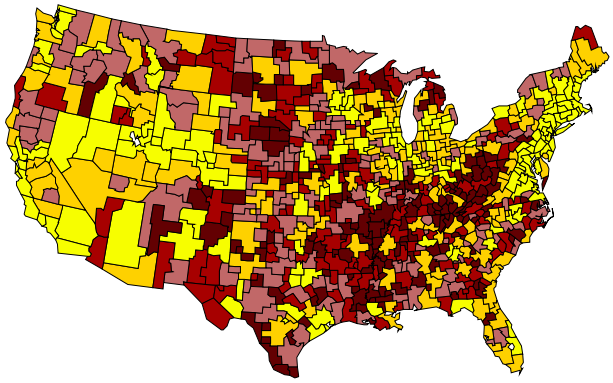


Model 1
22.759 - 23.91
22.485 - 22.759
22.17 - 22.485
21.665 - 22.17
18.825 - 21.665

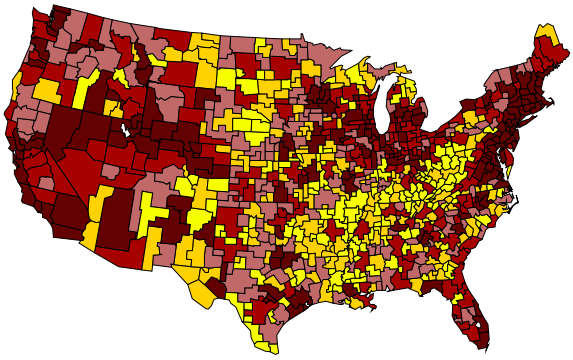


Model 1
62.709 - 67.227
61.911 - 62.709
61.416 - 61.911
60.983 - 61.416
59.177 - 60.983

Age 70



Model 1
33.652 - 34.667
33.404 - 33.652
33.115 - 33.404
32.647 - 33.115
29.855 - 32.647



Model 1
33.498 - 37.81
32.784 - 33.498
32.344 - 32.784
31.965 - 32.344
30.425 - 31.965

Fig 2.5: Maps of the Proportions of Deaths (10-2)

2.1.3 Concluding Remarks

We fitted a simple approximate model to the p_{ijk} that does not account for heterogeneity among health service areas. As expected, this approximate model does not appear to fit the data very well. Our mapping suggests that the proportion of death follow the same pattern across age classes and specific types of cancer, of course, with different intensities which give us the same pattern for the all cancer map and the specific disease maps.

The objective in the next section is to provide distinct patterns for each disease across age classes. To do so we investigate the fit of another approximate model which will serve as an alternative and improved model to the previous one by obtaining slightly smoother estimates and maps with age and disease specific patterns.

2.2 Second Approximate Model for the p_{ijk}'s

2.2.1 Model Description

We only modify the model for the p_{ijk}'s and use the same model for the λ_{ij} as the one we proposed in Section 2.1. We retain the 7 age classes.

We looked at the multinomial logit fixed effects model

$$y_{ijk} = \log\left(\frac{p_{ijk}}{p_{ij4}}\right) = \underline{Z}_i'\underline{\alpha} + \theta_{jk} + e_{ijk},$$

where $i = 1, \ldots, N,\ \ j = 1, \ldots, c,\ \ k = 1, 2, 3$, and $\underline{e}_{ij} \overset{ind}{\sim} N\big(\underline{0}, \sigma_j^2 \underset{3*3}{w_{ij}}\big)$.

$\underline{Z}_i'$ is defined as in Section 2.2 but without an intercept, $\sigma_j^{-2} \sim \Gamma(\frac{a}{2}, \frac{b}{2})$, $a = b = 0.002$ for $j = 1, \ldots, c$, and there are no constraints on the θ_{jk}.

Then $\underline{y}_{ij} \overset{ind}{\sim} N\big(\underline{Z}_i'\underline{\alpha} + \underline{\theta}_j, \sigma_j^2 W_{ij}^{-1}\big)$.

Letting $N = 798$, $c = 7$ and $W_{ij}^{-1} = \Omega_{ij}$, the joint posterior density obtained by assuming flat priors for all location parameters is

$$\begin{aligned} p(\underline{\alpha}, \underline{\theta}_j, \sigma_j^2 | \underline{d}) \;\; &\propto \;\; \prod_{i=1}^{N}\prod_{j=1}^{c}\left\{\left(\frac{1}{\sigma_j^2}\right)^{3/2} e^{-\frac{1}{2\sigma_j^2}\,(\underline{y}_{ij}-(\underline{Z}_i'\underline{\alpha}+\underline{\theta}_j))'\, W_{ij}^{-1}\,(\underline{y}_{ij}-(\underline{Z}_i'\underline{\alpha}+\underline{\theta}_j))}\right\} \\ &\times \;\; \prod_{j=1}^{c}\left(\frac{1}{\sigma_j^2}\right)^{a/2+1} e^{-\frac{b}{2\sigma_j^2}}. \end{aligned}$$

Then we can deduce the posterior distribution

$$\sigma_j^{-2} | \underline{\alpha}, \underline{\theta}_j, \underline{d} \;\sim\; \Gamma\Bigg(\;\frac{3N+a}{2}\;,\;\frac{b + \sum_i (\underline{y}_{ij} - (\underline{Z}_i'\underline{\alpha} + \underline{\theta}_j))'\;\Omega_{ij}\;(\underline{y}_{ij} - (\underline{Z}_i'\underline{\alpha} + \underline{\theta}_j))}{2}\Bigg). \tag{2.3}$$

Since the conditional posterior densities of $\underline{\alpha}$ and $\underline{\theta}_j$ are difficult to identify, we use the second order Taylor's series approximation. We Δ denote the logarithm of the conditional posterior density for each parameter such as for example for $\underline{\alpha}$

$$\Delta(\underline{\alpha}) \propto \sum_{i,j}\;\frac{1}{\sigma_j^2}\sum_{k_1,k_2=1}^{3}(y_{ijk_1} - (\underline{Z}_i'\underline{\alpha} + \theta_{jk_1}))\;\Omega_{ijk_1k_2}\;(y_{ijk_2} - (\underline{Z}_i'\underline{\alpha} + \theta_{jk_2}))$$

Then

$$\begin{aligned} \frac{d\Delta(\underline{\alpha})}{d\underline{\alpha}} \;\; &= \;\; -\sum_{i,j}\;\frac{1}{\sigma_j^2}\sum_{k_1,k_2=1}^{3}\;\Omega_{ijk_1k_2}\;(y_{ijk_2} - (\underline{Z}_i'\underline{\alpha} + \theta_{jk_2}))\underline{Z}_i \\ \frac{d^2\Delta(\underline{\alpha})}{d\underline{\alpha}^2} \;\; &= \;\; \sum_{i,j}\;\frac{1}{\sigma_j^2}\sum_{k_1,k_2=1}^{3}\;\Omega_{ijk_1k_2}\;\underline{Z}_i\;\underline{Z}_i' \end{aligned}$$

Therefore, we approximate the distribution of $\underline{\alpha}$ by

$$\underline{\alpha} | \underline{\theta}_j, \sigma_j^{-2} \overset{ind}{\sim} N\Bigg\{\Bigg(\sum_{i,j}\;\frac{1}{\sigma_j^2}\sum_{k_1,k_2=1}^{3}\;\Omega_{ijk_1k_2}\;\underline{Z}_i\;\underline{Z}_i'\Bigg)^{-1}$$

$$\Bigg(\sum_{i,j}\;\frac{1}{\sigma_j^2}\sum_{k_1,k_2=1}^{3}\;\Omega_{ijk_1k_2}\;(y_{ijk_2} - \theta_{jk_2})\underline{Z}_i\Bigg)\;,\;\Bigg(\sum_{i,j}\;\frac{1}{\sigma_j^2}\sum_{k_1,k_2=1}^{3}\;\Omega_{ijk_1k_2}\;\underline{Z}_i\;\underline{Z}_i'\Bigg)^{-1}\Bigg\}. \tag{2.4}$$

By the same methodology, we approximate the distribution of $\underline{\theta}_j$ by

$$\underline{\theta}_j|\underline{\alpha}, \sigma_j^{-2} \overset{ind}{\sim} N\left\{\left(\sum_i \Omega_{ij}\right)^{-1}\left(\sum_i \Omega_{ij}\ (\underline{y}_{ij} - \underline{Z}_i'\underline{\alpha})\right),\ \sigma_j^2\left(\sum_i \Omega_{ij}\right)^{-1}\right\} \quad (2.5)$$

We use the Gibbs sampler to generate 1000 estimates of $\underline{\alpha}, \underline{\theta}$ drawing from (2.3), (2.4) and (2.5) in turn; the convergence was rapid. In fact, we dropped out the first 100 and took the next 1000 iterates to make inference.

Finally, we deduce the corresponding 1000 p_{ijk} as explained in Appendix B from the Gibbs estimates so we have an estimate of the distribution of the p_{ijk} each based on 1000 samples. The idea is to have smooth estimates of the p_{ijk}. This is a slightly refined procedure relative to the first one presented in the previous section.

We could have separated out θ_{jk} into its components as in the first approximation but we simplify the computations somewhat.

2.2.2 Estimates

We look at two cases. In the first case the reference parameter is other cancer and the second case the reference parameter is prostate cancer.

Reference Parameter is Other Cancer

Tables 2.3 and 2.4 present the means over the 1000 simulations of the parameter estimates. The estimates follow the patterns of the observed mortality data. All the parameters are significant except poverty.

In Table 2.5 we present $E(.|\underline{d})$ and $SD(.|\underline{d})$ for the age specific and age adjusted (age class 8) rates average over HSAs. They look similar to the observed ones except for colon cancer for which the estimated rates are underestimated after age class 2 and

Covariates	Estimates			
	Mean	Std	Interval	Ratio
% Income	-0.112	0.018	[-0.124, -0.100]	-6.22
% College	-0.007	0.003	[-0.009, -0.006]	-2.33
% Poverty	-0.004	0.0027	[-0.006, -0.003]	-1.48
Epapm25	-4.709	0.988	[-5.387 ,-4.034]	-4.77
EpaSO2	0.7217	0.251	[0.553, 0.883]	2.87

Table 2.3: Mean and Standard Deviation of the Estimates of α

	Prostate		Colon		Lung	
Age Class	Mean	Std	Mean	Std	Mean	Std
1	-13.713	0.350	-2.780	0.081	-3.087	0.078
2	-15.624	0.325	-1.761	0.047	-0.740	0.038
3	-4.721	0.107	-1.857	0.046	0.127	0.037
4	-2.894	0.047	-1.749	0.042	0.478	0.035
5	-1.636	0.038	-1.529	0.040	0.499	0.035
6	-0.659	0.035	-1.250	0.036	0.220	0.034
7	0.037	0.035	-0.948	0.039	-0.484	0.037

Table 2.4: Mean and Standard Deviation of the Estimates of θ_{jk}

for prostate cancer for which the estimated rates are underestimated until age class 6.

	All Cancer		Prostate		Colon		Lung	
Age Class	Mean	Std	Mean	Std	Mean	Std	Mean	Std
1	6.078	1.160	.0000	.0000	.2745	.0534	.2020	.0393
2	37.51	5.198	.0000	.0000	3.367	.4914	9.342	1.364
3	152.8	20.69	.5323	.0748	9.292	1.305	67.56	9.488
4	512.3	62.19	9.123	1.141	28.65	3.582	265.6	33.21
5	1069	149.2	62.64	8.991	69.68	10.00	529.6	76.01
6	1887	187.6	294.1	29.68	162.9	16.44	708.2	71.48
7	2723	260.6	853.5	82.02	318.7	30.63	506.7	48.69
8	159.9	17.42	11.28	1.105	11.10	1.215	70.01	8.103

Table 2.5: Mean and Standard Deviation per 100,000 population of the Death Rates (λ_{ijk}) over HSAs

In Figure 2.6 we present the distribution of the proportions of death (p_{ijk}) by disease. The patterns of the box plots across the age classes look similar to the one obtained from the observed proportions in Figure 1.4 for prostate, lung and other cancers but smoothed. The patterns across age classes and the estimated proportions are similar.

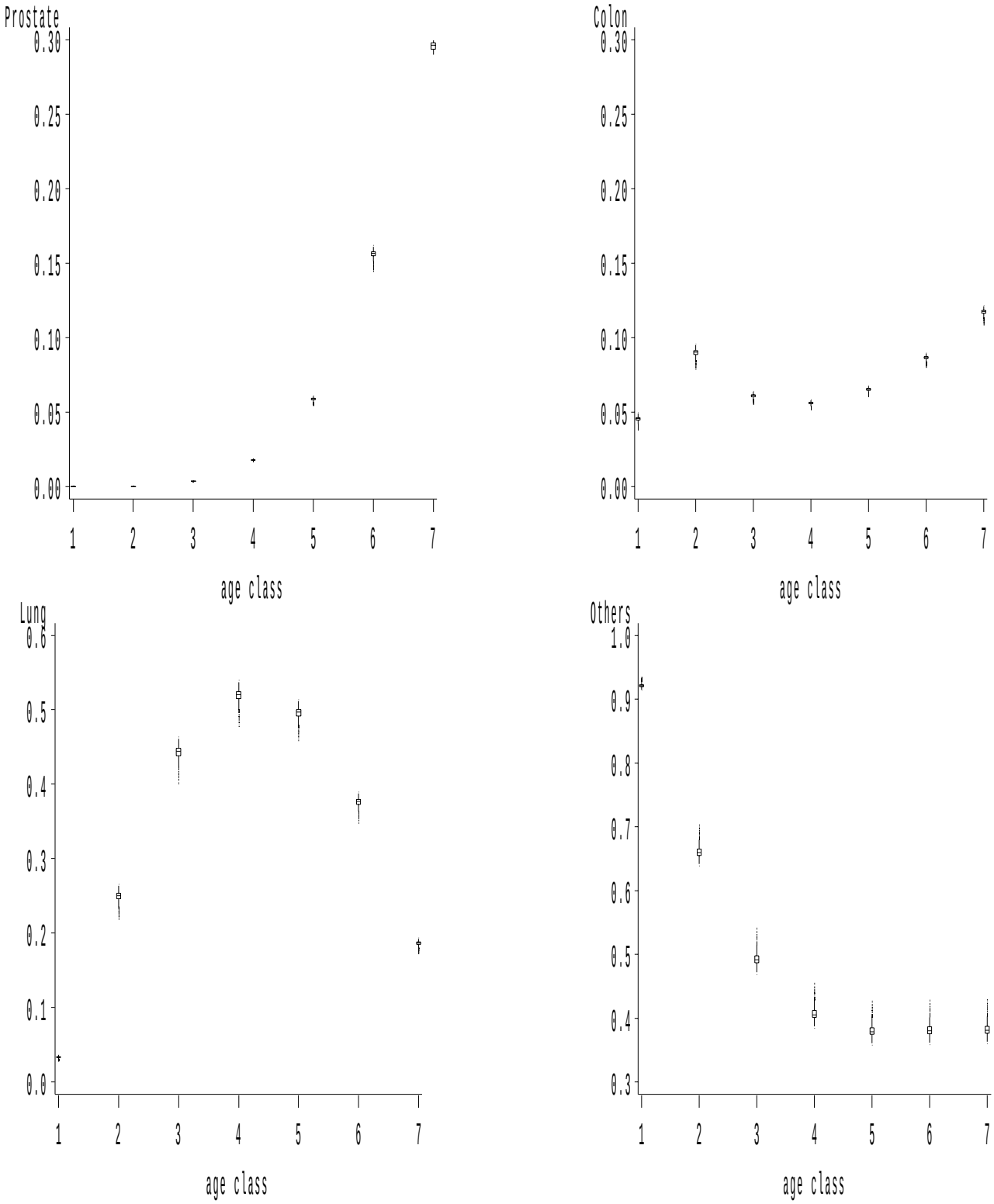


Figure 2.6: Box Plots of the Estimated Proportions of Death by Disease

Reference Parameter is Prostate Cancer

Tables 2.6 and 2.7 present the means over the 1000 simulations of the parameter estimates. The estimates follow the patterns of the observed mortality data. All the parameters are significant except epapm25.

Covariates	Estimates			
	Mean	Std	Interval	Ratio
% Income	0.232	0.039	[.206,.258]	5.95
% College	-0.037	0.006	[-.041,-.033]	-6.17
% Poverty	0.008	0.004	[.005,.011]	2
Epapm25	2.987	2.098	[1.59,4.39]	1.42
EpaSO2	1.188	0.538	[.829,1.55]	2.21

Table 2.6: Mean and Standard Deviation of the Estimates of α

	Prostate		Colon		Lung	
Age Class	Mean	Std	Mean	Std	Mean	Std
1	8.938	0.160	10.392	1.393	17.511	1.364
2	9.431	0.190	12.234	0.676	15.569	0.411
3	2.253	0.177	5.218	0.443	6.514	0.213
4	-0.186	0.088	2.637	0.146	2.886	0.091
5	-0.842	0.073	1.588	0.086	1.760	0.075
6	-1.110	0.073	0.591	0.079	1.128	0.074
7	-1.173	0.074	-0.597	0.088	0.774	0.082

Table 2.7: Mean and Standard Deviation of the Estimates of θ_{jk}

In Table 2.8 we present $E(.|\underline{d})$ and $SD(.|\underline{d})$ for the age specific and age adjusted (age class 8) rates average over HSAs. They look similar to the observed rates except for colon and prostate cancer for which the estimated rates are underestimated.

In Figure 2.7 we present the distribution of the proportions of deaths by disease. The patterns of the box plots across age classes and the estimated proportions of death little bit different from the one obtained from the observed proportions in Figure 1.4

	All Cancer		Prostate		Colon		Lung	
Age Class	Mean	Std	Mean	Std	Mean	Std	Mean	Std
1	6.078	1.160	.0000	.0000	.0029	.0006	.0345	.0066
2	37.51	5.198	.0000	.0000	.0858	.0119	1.686	.2336
3	152.8	20.69	.1399	.0183	1.708	.2314	34.04	4.610
4	512.3	62.19	12.20	1.421	12.71	1.547	213.7	26.02
5	1069	149.2	71.54	9.612	38.59	5.422	438.5	61.62
6	1887	187.6	250.1	23.95	103.2	10.49	565.7	57.49
7	2723	260.6	568.2	54.19	220.2	21.71	391.9	38.63
8	159.9	17.42	10.36	.9589	5.500	.5681	53.54	6.003

Table 2.8: Mean and Standard Deviation per 100,000 population of the Death Rates (λ_{ijk}) over HSAs

for prostate, lung and other cancers. Indeed the proportions of death for prostate and colon cancer are underestimated.

The box plots obtained with other cancer as the reference match the observed ones better.

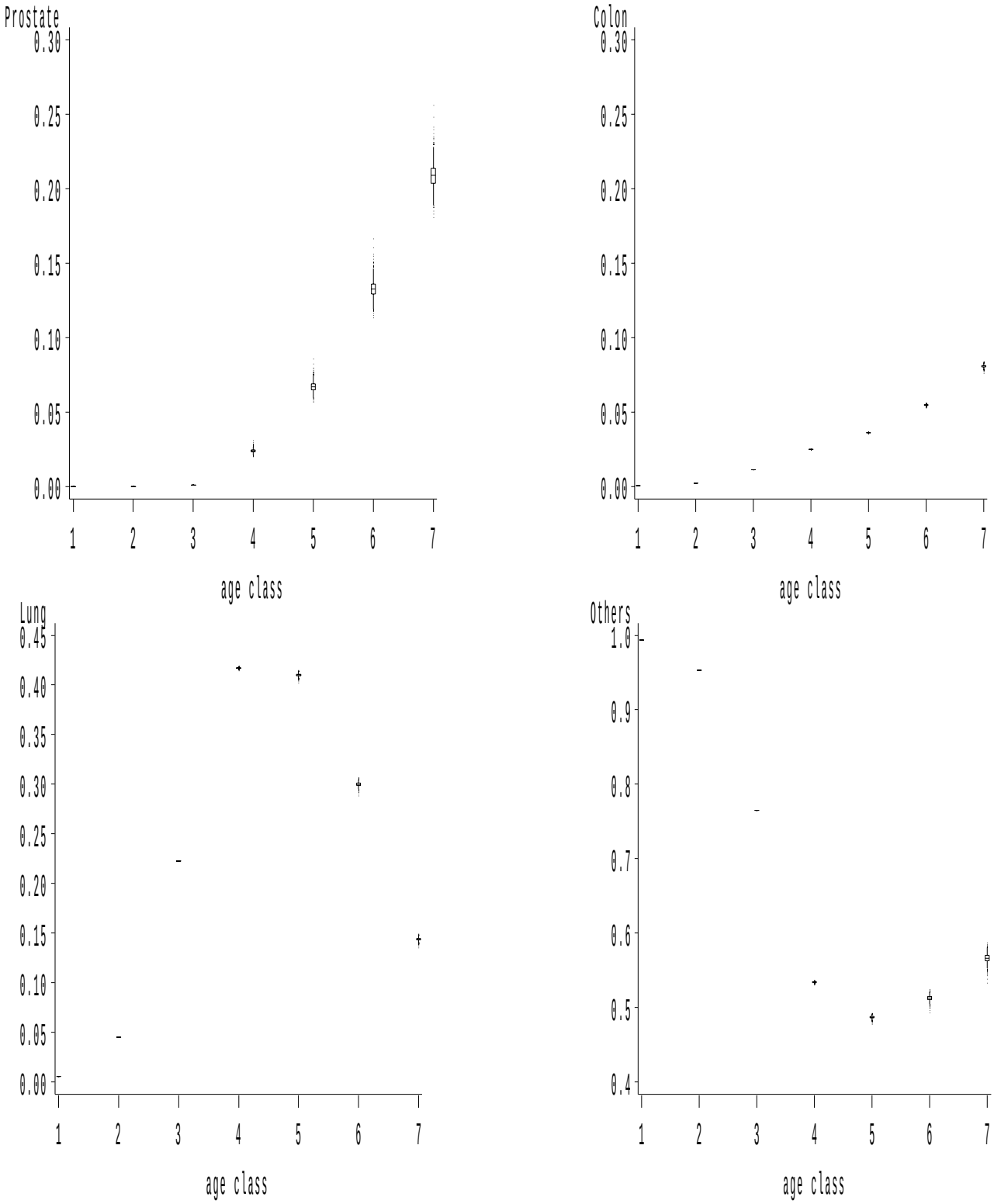


Figure 2.7: Box Plots of the Estimated Proportions of Deaths by Disease

2.2.3 Maps

Since the results are similar but match the observed mortality rates better with other cancer as the reference parameter, we only drew and analyzed the corresponding maps.

In Figures 2.8 and 2.9 we present the maps for age specific (40, 70) and age adjusted mortality rates by disease. The same pattern of concentration of high mortality rates around the Appalachian region (Mississippi to West Virginia) is observed for each age class across disease. Only a slight improvement is observed compared to the first approximate model fitted in Section 2.2.

In Figures 2.10 and 2.11 we present the maps for age specific 40 and 70 of the proportions of deaths by type of cancer. The map for prostate cancer at age 40 clearly show that the model encounters major difficulties. The remaining maps show some improvement from the first model since they match the maps of the observed proportions presented in chapter 1 better.

Figure 2.8: Maps of the Estimated Death Rates by Disease

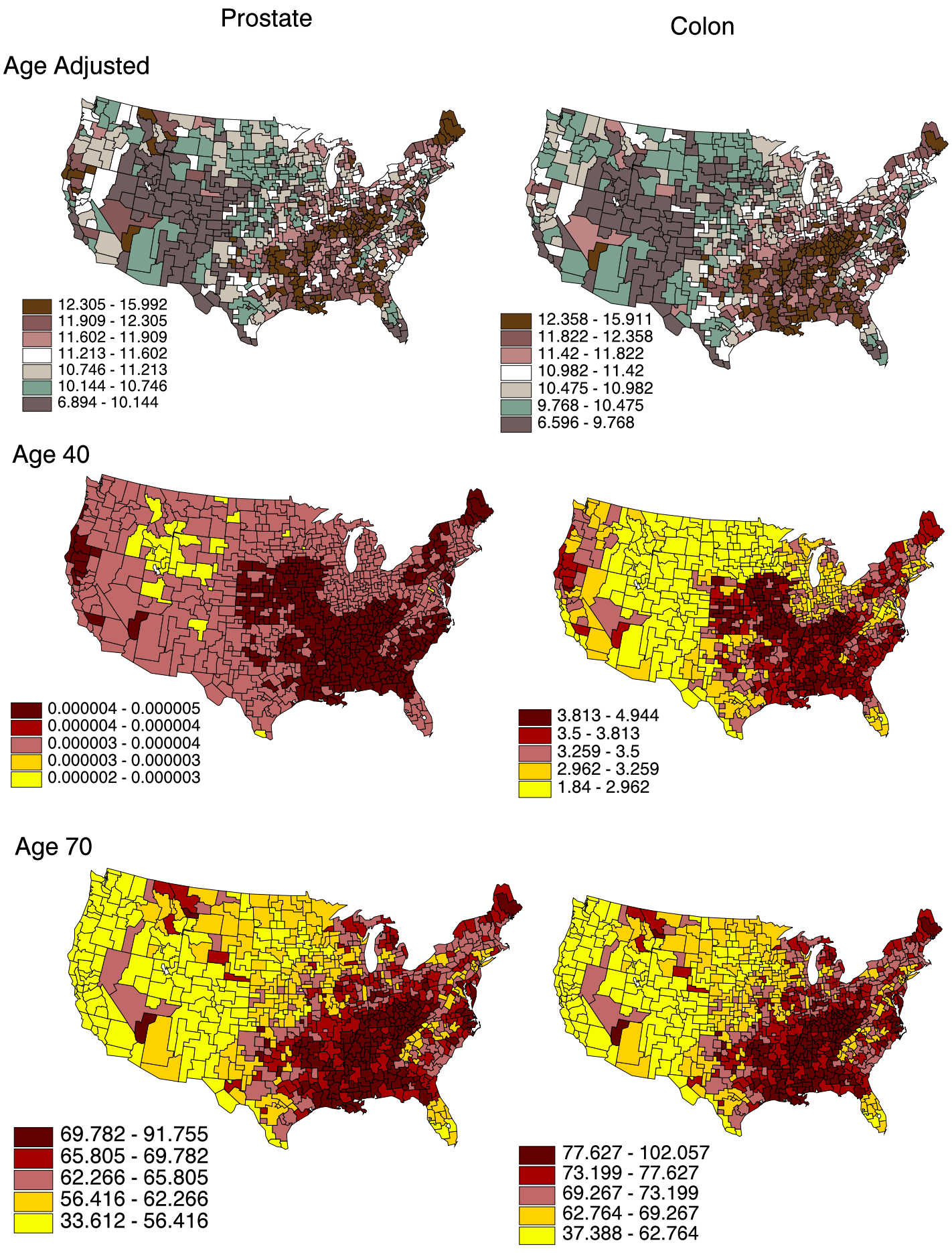
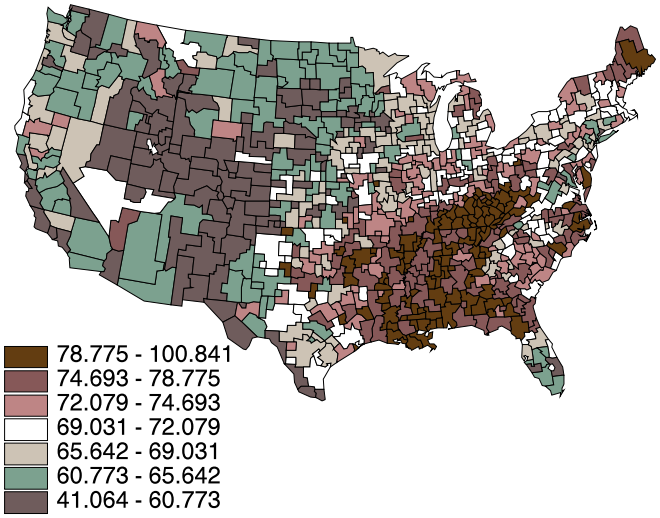


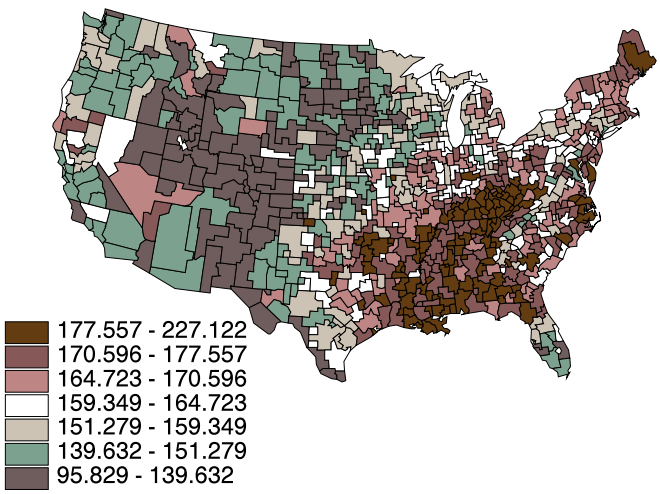
Fig. 2.9: Maps of the Estimated Death Rates by Type of Cancer

Lung

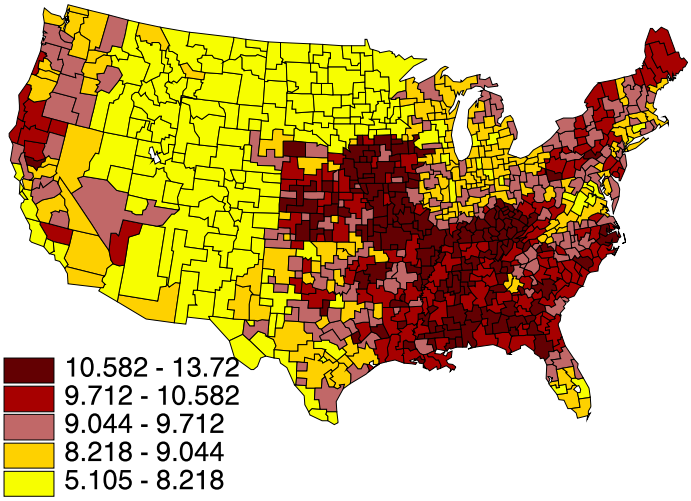
All Cancer

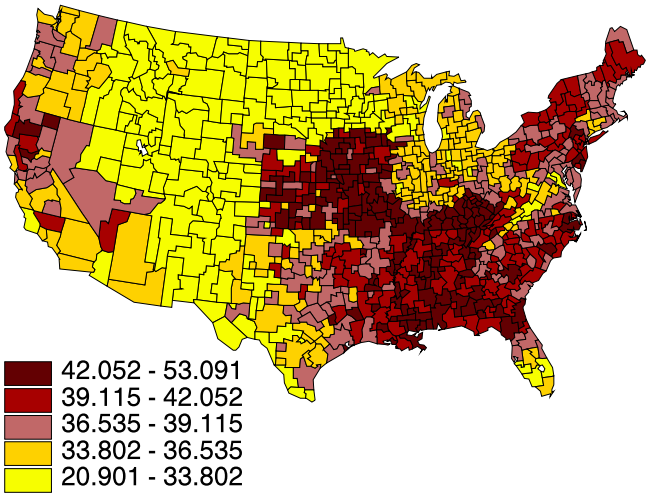
Age Adjusted



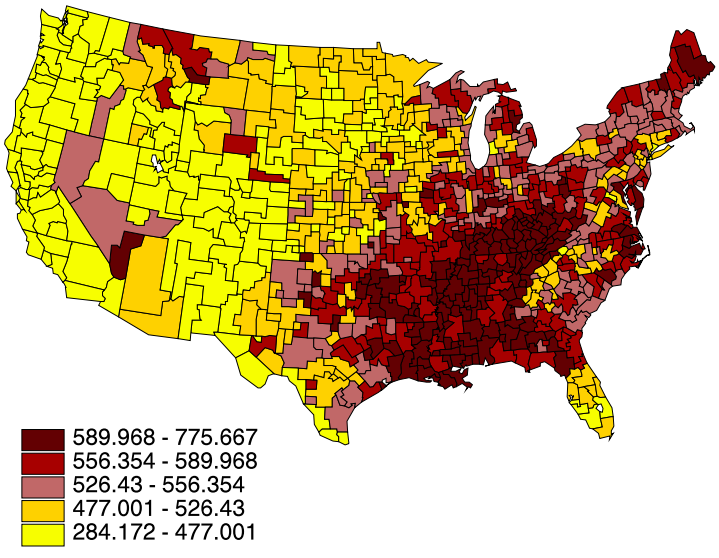


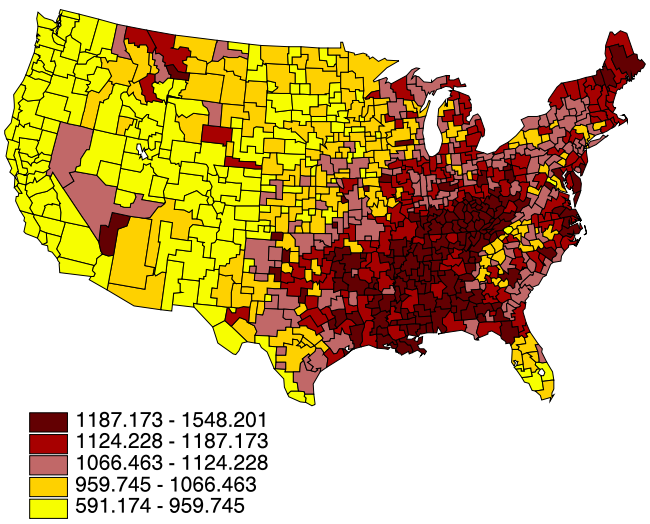
Age 40





Age 70





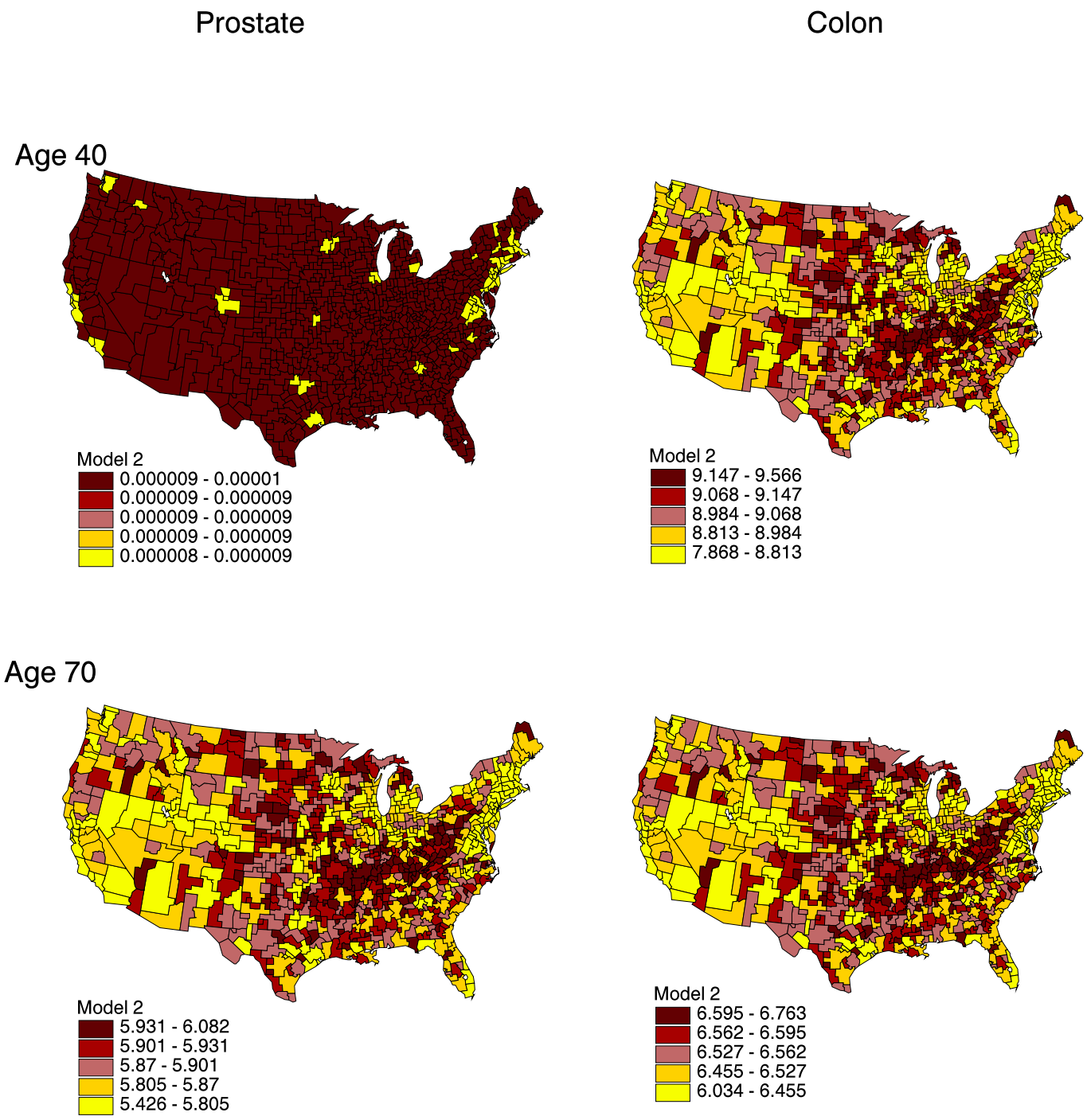
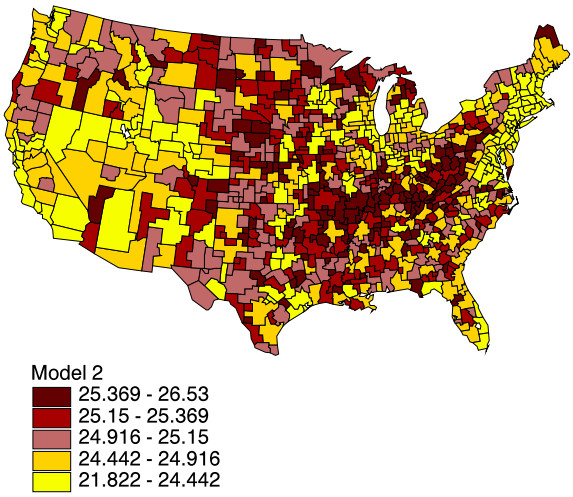


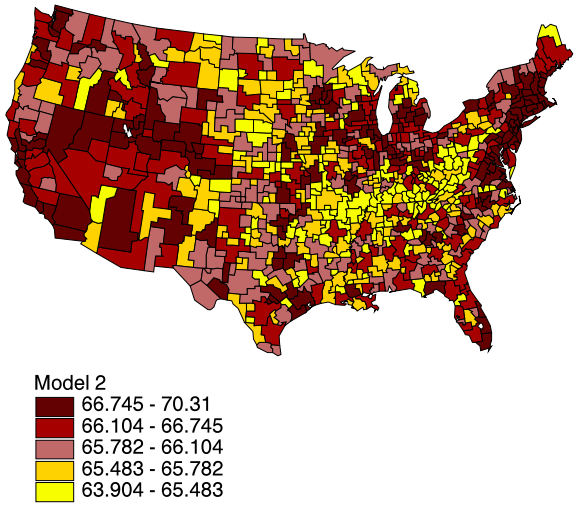
Fig 2.10: Maps of the Proportions of Deaths (10-2)

Lung

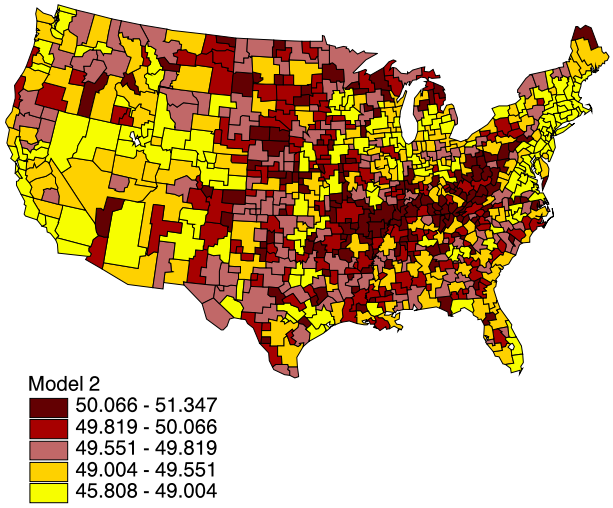
Other

Age 40





Age 70



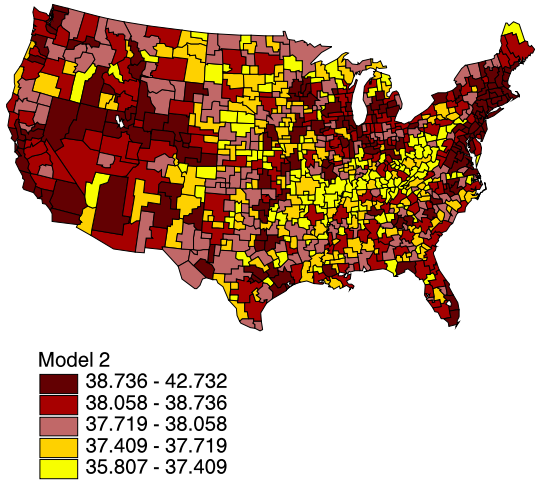


Fig 2.11: Maps of the Proportions of Deaths (10-2)

2.3 Concluding Remarks

In this section we introduced two simple approximate models for the p_{ijk} that seem to account well for heterogeneity among Health Service Areas for each disease. The maps have the same pattern for each disease.

Only slight improvements were observed with the second model. The estimates are still very smoothed and our mapping still suggests that the proportions for each disease category do not change among HSA. Besides the distribution of the proportions of death by disease obtained from the second approximation are closer to the observed ones.

Since the data are very sparse for prostate cancer, we fitted the model twice, once with reference parameter others cancer and once with prostate cancer. The results are slightly closer to the observed data with other cancer as the reference.

In Chapter 3, we follow up further by first fitting the model used previously for the λ_{ij} to the λ_{ijk} and then fit an improve model to the p_{ijk}.

Chapter 3

An Exact Model

The objective in this chapter is to investigate if there are different map patterns for each type of disease, and to obtain improved precision.

The models we studied in the previous chapters did not fit the data well because they produce one map pattern among the different types of cancer. In this chapter, we first focus on a model for the λ_{ijk} which is a generalization of the model presented in Chapter 2 for the λ_{ij} to each disease, and then we deduce the proportions of death by disease p_{ijk}.

Note, in particularly, in this chapter we use all 10 age classes.

3.1 Model for the λ_{ijk}

The basis model generalized from (1.7) is as follows

$$\log \lambda_{ijk} = \underline{x}_j^{'} \, \underline{\beta_k} + \nu_{ik} + \delta_{jk}$$

where $\underline{x}_j^{'} = \big(1, decade_j, (decade_j)^2, (decade_j)^3, max\{0, (decade_j - knot)^3\}\big)$ with $decade_1 = 0.25$, $decade_j = j - 1$ for $j = 2, \ldots, 10$ and $k = 1, \ldots, 4$ for each disease prostate, colon, lung and others cancer respectively.

It is assumed that $\nu_{ik}|\sigma_{1k}^2 \overset{iid}{\sim} N\big(0, \sigma_{1k}^2\big)$, $\delta_{jk}|\sigma_{2k}^2 \overset{iid}{\sim} N\big(0, \sigma_{2k}^2\big)$, and the value of the knot that maximizes the likelihood of U.S. marginal data is 6 for "all cancer".

Here, $p(\underline{\beta}) = 1$ and $\sigma_{1k}^{-2}, \sigma_{2k}^{-2} \sim \Gamma(\frac{a}{2}, \frac{b}{2})$ where $a = b = 0.002$ to obtain a proper diffuse prior.

The Metropolis-Hastings algorithm was used based in a manner similar to that in Appendix A but applied to each disease.

By trial and error, we chose the tuning constants between 6 and 10 (see Appendix A).

The computations were done by region. We ran $11,000$ iterates, "burning in" 1000 and choosing every tenth to obtain 1000 iterates which we used for output analyzes.

3.2 Assessing the Model Fit

We have used two different measures to assess the model.

The first measure that we used to assess the model is the posterior predictive p-value; i.e.,

$$\Pr\{T(\underline{d}^{new}, \lambda) \geq T((d^{obs}, \underline{\lambda}) \mid \underline{d}^{obs}\}. \tag{3.1}$$

Very small or very large values of (3.1) are sometimes used to discredit a model (Gelman et al. 1995, Chapter 6). A model is considered acceptable if the p-value is between 0.05 and 0.95. We have used three checking functions, $T(\underline{d}^{new}, \underline{\lambda})$, analogous to the three discrepancy measures, $P(\underline{d}^{obs}, \underline{d}^{new})$:

1. Chi-square

$$\sum_i \sum_j (d_{ijk} - n_{ij}\lambda_{ijk})^2 / n_{ij}\lambda_{ijk}.$$

2. Rank-based

$$\sqrt{12}\sum_i\sum_j\{c_{ijk}/(a+1)-0.5\}(d_{ijk}-n_{ij}\lambda_{ijk})$$

where $a = 10$ and $c_{ijk} = \text{rank}(d_{ijk} - n_{ij}\lambda_{ijk})$.

3. Poisson-based

$$2\sum_i\sum_j\left\{(d_{ijk}+0.5)\ln\left(\frac{d_{ijk}+0.5}{n_{ij}\lambda_{ijk}+0.5}\right)-(d_{ijk}-n_{ij}\lambda_{ijk})\right\}.$$

Region	Prostate			Colon		
	Chi-Square	Poisson-based	Rank-based	Chi-Square	Poisson-based	Rank-based
1	0.157	0.185	0.125	0.418	0.116	0.126
2	0.920	0.642	0.237	0.519	0.149	0.055
3	0.935	0.745	0.406	0.271	0.000	0.001
4	0.222	0.051	0.060	0.458	0.005	0.072
5	0.928	0.540	0.418	0.884	0.318	0.328
6	0.791	0.532	0.283	0.305	0.000	0.005
7	0.729	0.099	0.036	0.220	0.282	0.461
8	0.923	0.721	0.637	0.615	0.876	0.568
9	0.348	0.227	0.188	0.305	0.002	0.022
10	0.426	0.117	0.165	0.373	0.372	0.388
11	0.948	0.454	0.393	0.275	0.676	0.779
12	0.874	0.395	0.103	0.187	0.002	0.001

Table 3.1: Posterior Predictive P-values by Region for Prostate and Colon Cancer

In Table 3.1 and 3.2 we present the p-values of the three checking functions of the model by region for each disease. Most of the chi-square p-values lie between 0.05 and 0.95 across region for prostate and colon cancer but not for lung and other cancer. The fit for lung cancer seems unreasonable based on this measure. These p-values are very sensitive to outliers and an extreme outlier could force them to zero.

The second method of evaluating the model is to use a cross-validation. Let $\underline{d}_{(ijk)}$ denote the set of all d's *except* for (ijk). Then define the cross-validation residual as $a_{ijk} = r_{ijk} - E(r_{ijk} \mid \underline{d}_{(ijk)})$, and the standardized cross-validation residual as

$$\text{DRES}_{ijk} = \frac{r_{ijk} - E(r_{ijk} \mid \underline{d}_{(ijk)})}{SD(r_{ijk} \mid \underline{d}_{(ijk)})}. \tag{3.2}$$

Region	Lung			Others		
	Chi-Square	Poisson-based	Rank-based	Chi-Square	Poisson-based	Rank-based
1	0.120	0.044	0.018	0.575	0.628	0.303
2	0.238	0.000	0.002	0.000	0.000	0.000
3	0.083	0.000	0.004	0.000	0.000	0.020
4	0.004	0.000	0.000	0.000	0.000	0.000
5	0.134	0.000	0.000	0.194	0.061	0.121
6	0.008	0.000	0.000	0.428	0.515	0.611
7	0.312	0.255	0.019	0.009	0.024	0.205
8	0.103	0.000	0.000	0.011	0.045	0.009
9	0.097	0.000	0.000	0.068	0.037	0.029
10	0.062	0.000	0.000	0.093	0.066	0.030
11	0.258	0.024	0.058	0.194	0.356	0.285
12	0.218	0.022	0.008	0.002	0.000	0.046

Table 3.2: Posterior Predictive P-values by Region For Lung and Other Cancer

That is, the (ijk)-th observed r_{ijk} is "held out" and compared with its point estimator, $E(r_{ijk} \mid \underline{d}_{(ijk)})$, which is evaluated *without* using the observed d_{ijk}. We employ the cross-validation residuals and standardized residuals as absolute measures of the concordance of the data with a proposed model. To summarize we count (a) the number of (ijk) such that $\mid \text{DRES}_{ijk} \mid \geq q$ which we call "# outliers" and (b) the number of HSAs such that $\mid \text{DRES}_{ijk} \mid \geq q$ for at least one j, which we call "# HSAs". In Table 3.3 we present the number of HSAs with outliers ≥ 3 and 4 by region. (Table 1.4 contains the number of HSAs per region). As can be seen regions 4 and 9 have the most number of HSAs with residuals greater than 3 and 4. It indicates that regions 4 and 9 are not fitted well. Table 3.4 contains the number of HSAs with outliers ≥ 3 and 4 by age class. The model does not seem to fit the data very well at the age classes 5 and 9 for prostate cancer, 3 to 5 for colon cancer, all the age classes for lung cancer and all the age classes but the last one for other cancer.

The large number of HSAs with outliers comes from the fact that the data are very sparse across age class.

Finally we decide to assess the model by looking at the residual plots for all dis-

Region	Prostate		Colon		Lung		Others	
	#HSA≥3	#HSA≥4	#HSA≥3	#HSA≥4	#HSA≥3	#HSA≥4	#HSA≥3	#HSA≥4
1	1	1	5	2	17	0	2	0
2	4	2	5	1	11	5	7	5
3	3	1	8	2	18	2	10	2
4	38	20	10	2	28	2	11	2
5	5	3	8	1	14	0	5	0
6	12	6	19	4	44	1	7	1
7	2	1	6	2	4	3	10	3
8	5	4	13	5	21	7	13	7
9	10	6	21	8	43	3	13	3
10	5	2	5	3	24	1	4	1
11	0	0	8	3	4	3	7	3
12	4	0	12	7	8	2	8	2

Table 3.3: Number of HSAs with Absolute Values of Residuals ≥3 and ≥4 by Region

Age	Prostate		Colon		Lung		Others	
	#HSA≥3	#HSA≥4	#HSA≥3	#HSA≥4	#HSA≥3	#HSA≥4	#HSA≥3	#HSA≥4
1	0	0	0	0	8	7	17	8
2	1	1	3	1	6	5	13	3
3	7	5	20	8	18	13	13	3
4	3	3	32	11	17	7	9	3
5	24	16	20	4	12	6	8	3
6	9	6	12	1	34	5	14	4
7	8	2	8	3	42	19	8	1
8	9	2	8	5	67	36	5	1
9	23	10	10	5	27	10	9	3
10	5	1	7	2	10	4	1	0

Table 3.4: Number of HSAs with Absolute Values of Residuals ≥3 and ≥4 by Age Class

eases simultaneously.

In Figures 3.1 we present the plot of residual against standard deviation of residual, SD($r_{ijk}|d_{ijk}$) and the plot of a_{ijk} versus number of death. We provide bands of $|r_{ijk} - E(r_{ijk}|d_{(ijk)})| \leq 2\ SD(r_{ijk}|d_{ijk})$ which in fact gives it a funnel shape, where both sides of the funnel have some outliers. The other plot is residual a_{ijk} against number of death d_{ijk}. This is indicating that the model provides a good fit.

Figures 3.2 are the plots of residual a_{ijk} against age and region. The plot against age shows the increase in variation as the age increases, the highest being age class

10, i.e. 85 years and above. There is symmetry about zero as can be seen from the two types of plots, against age and region. These are further good indications of the fit of the model.

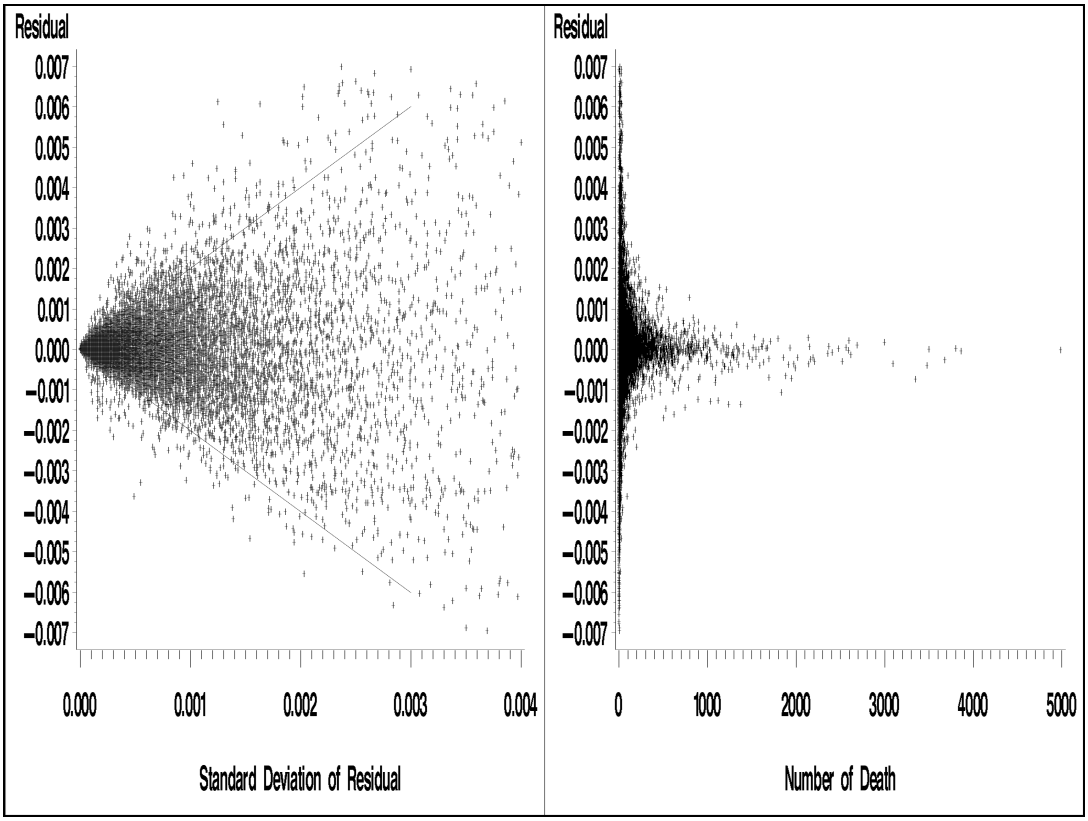


Figure 3.1: Plots of Residuals versus Standard Deviation of Residuals and Number of Deaths

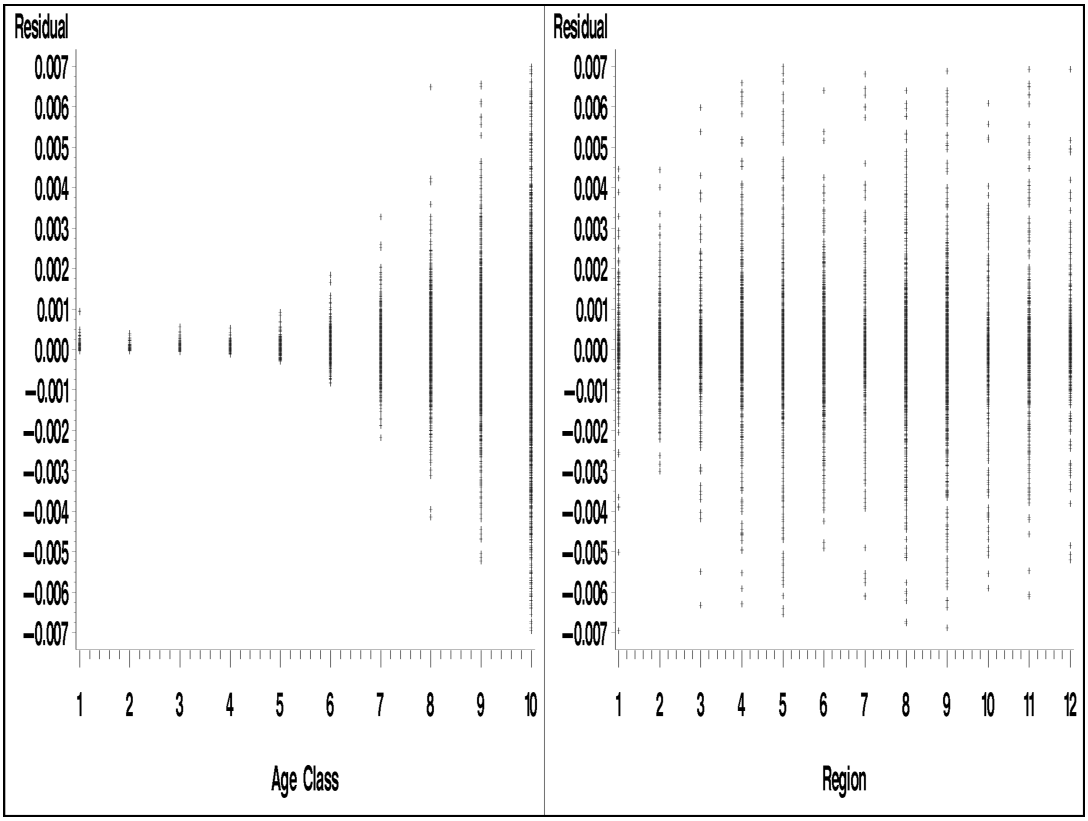


Figure 3.2: Plots of Residuals versus Age and Region

Figure 3.3 shows the plots of residual a_{ijk} and standardized residuals against the predictive rates, λ_{ijk} for the 4 diseases simultaneously. In a Normal distribution model, these plots should be a null plot, i.e. no pattern, but in our case it is a Poisson distribution model. The mean of the Poisson is directly proportional to the variance, hence the megaphone shape is expected. The interesting characteristic is its symmetry about zero.

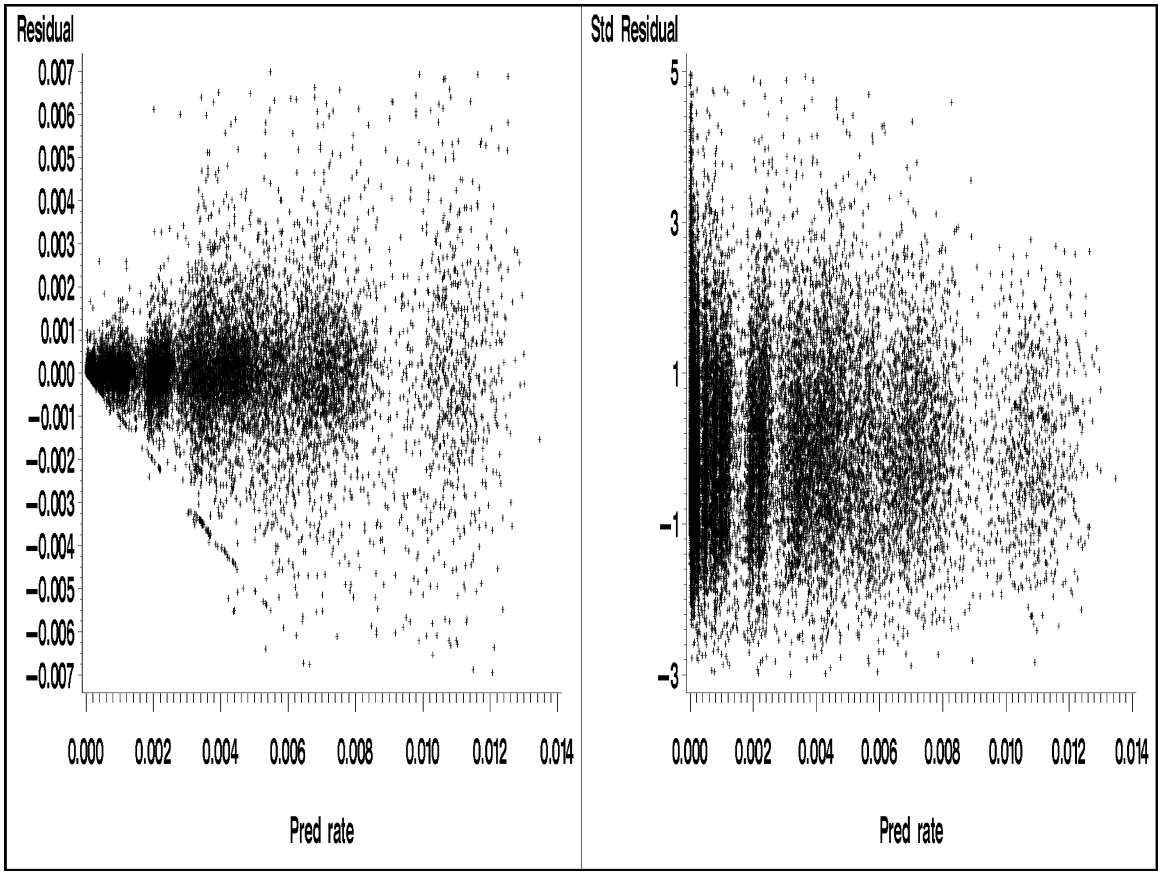


Figure 3.3: Plots of Residuals and Standardized Residuals versus Predicted Rates

The box plots of the standardized residuals versus age class, region and type of cancer presented in Figures 3.4, 3.5, and 3.6 do not show any departure from the symmetrical pattern about zero.

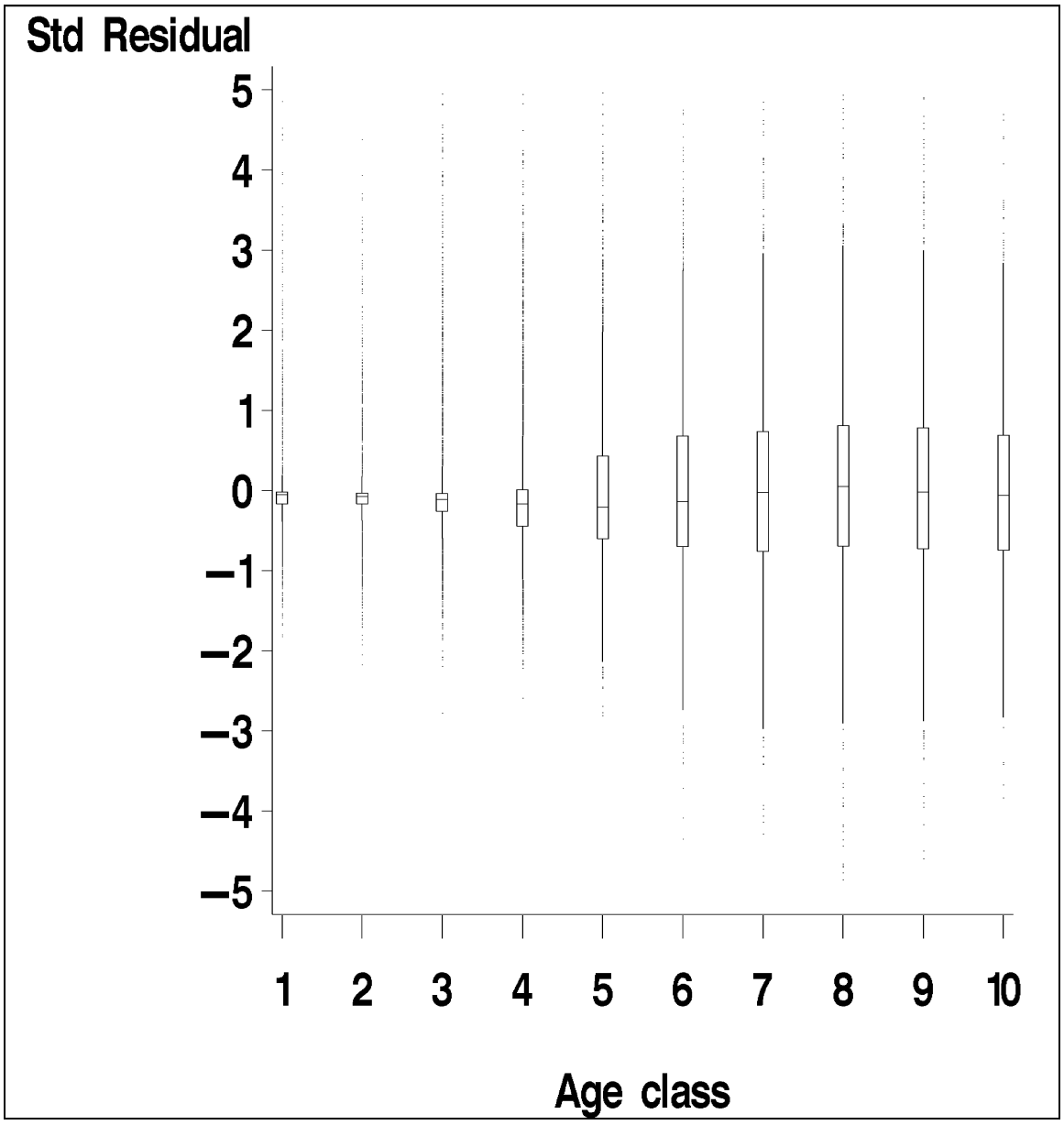


Figure 3.4: Box Plots of the Standardized Residuals versus Age Class

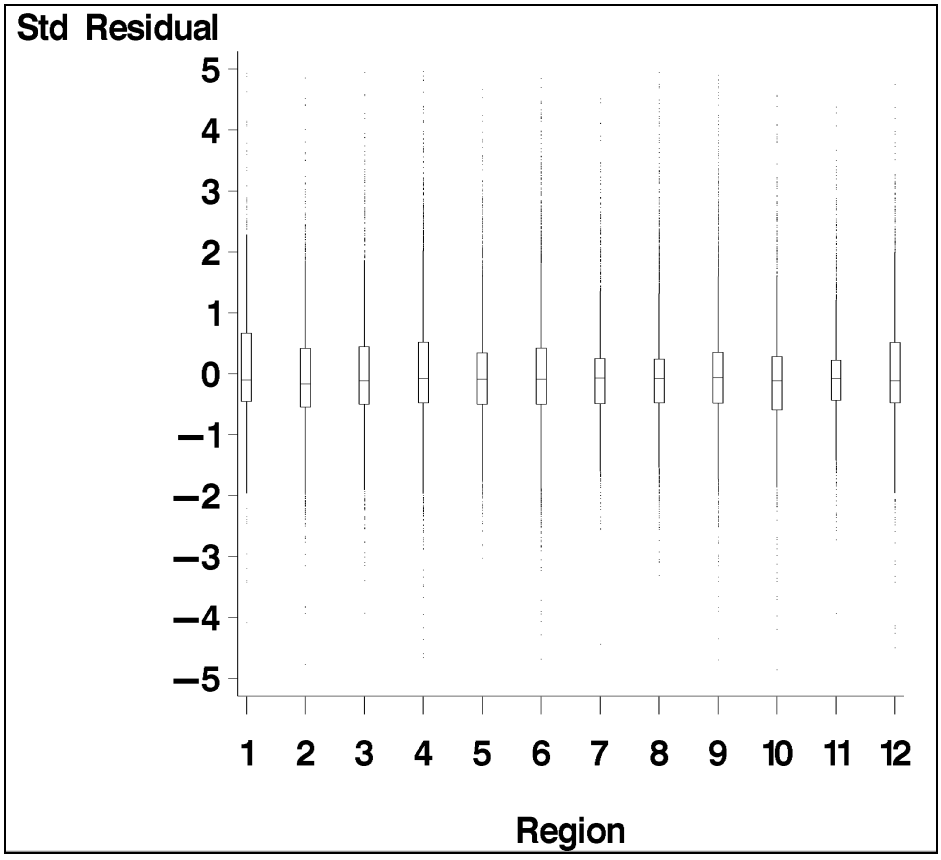


Figure 3.5: Box Plots of the Standardized Residuals versus Region

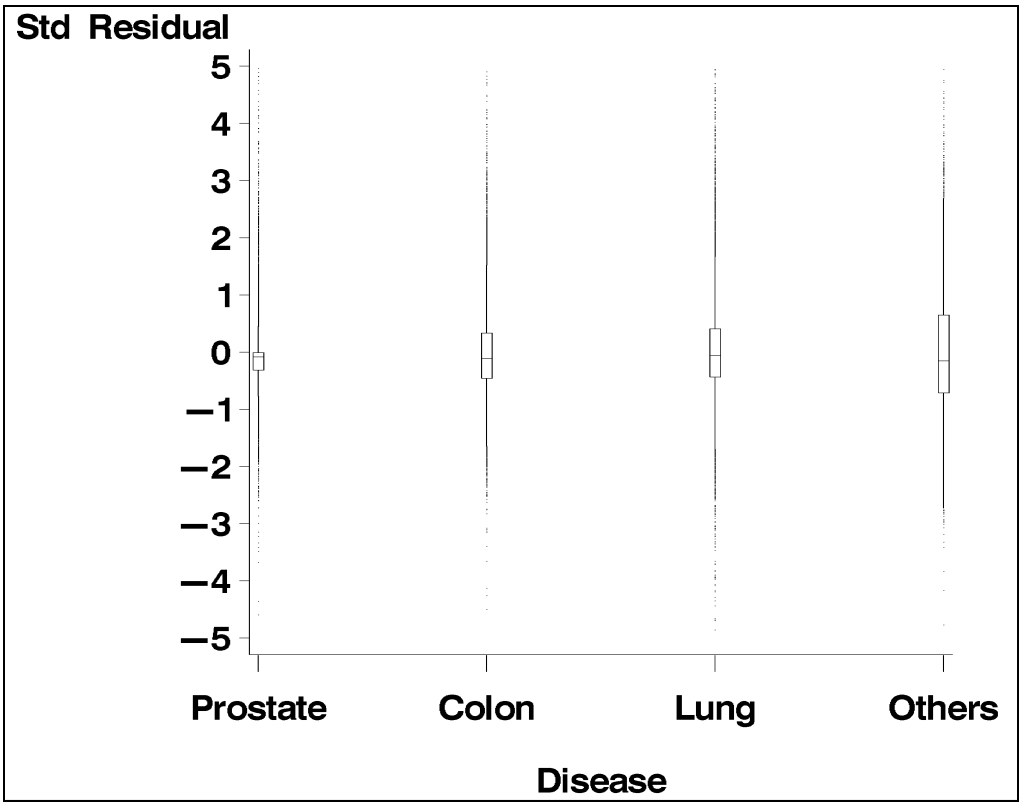


Figure 3.6: Box Plots of the Standardized Residuals versus Type of Cancer

3.3 Estimates and Maps for the Death Rates

Table 3.5 presents the mean and standard Deviation of the death rates, λ_{ijk}, over HSAs. The death rates increase with age class for each disease. They are very similar to the observed ones presented in Table 1.5.

	All Cancer		Prostate		Colon		Lung	
Age Class	Mean	Std	Mean	Std	Mean	Std	Mean	Std
1	3.686	.4384	.0548	.0655	.0297	.0300	.0097	.0164
2	3.551	.4011	.0121	.0071	.0345	.0171	.0117	.0119
3	5.584	.6812	.0114	.0058	.1127	.0413	.0609	.0291
4	11.68	1.375	.0262	.0202	.6111	.1648	.6845	.1791
5	36.20	5.016	.1576	.0694	3.140	.5620	7.519	1.876
6	151.7	21.71	2.263	.4328	13.94	1.662	55.86	13.02
7	509.6	64.12	22.68	3.269	47.51	5.335	213.8	42.93
8	1090	124.1	107.6	15.99	108.9	14.74	427.6	76.03
9	1870	188.1	326.8	47.80	214.1	27.43	572.9	86.58
10	2711	255.1	714.8	107.7	365.1	42.67	510.7	74.37
11	160.3	18.13	14.97	2.184	16.01	1.874	56.95	10.50

Table 3.5: Mean and Standard Deviation per 100,000 population of the Death Rates (λ_{ijk}) over HSAs

In Figure 3.7 and 3.8 we present the maps for age specific 40, 70 and age adjusted mortality rates by disease. Since the data are very sparse for age younger than 40 years for prostate and colon cancers, the corresponding maps have to be interpreted with caution. The maps show interesting hot spots for prostate cancer in the North West and North Central regions. For colon cancer, the high mortality rates are concentrated in the regions of the North East and east North Central. The high mortality rates for lung cancer and all cancers are both concentrated in the South East region for each age class.

Fig. 3.7: Maps of the Estimated Death Rates by Type of Cancer

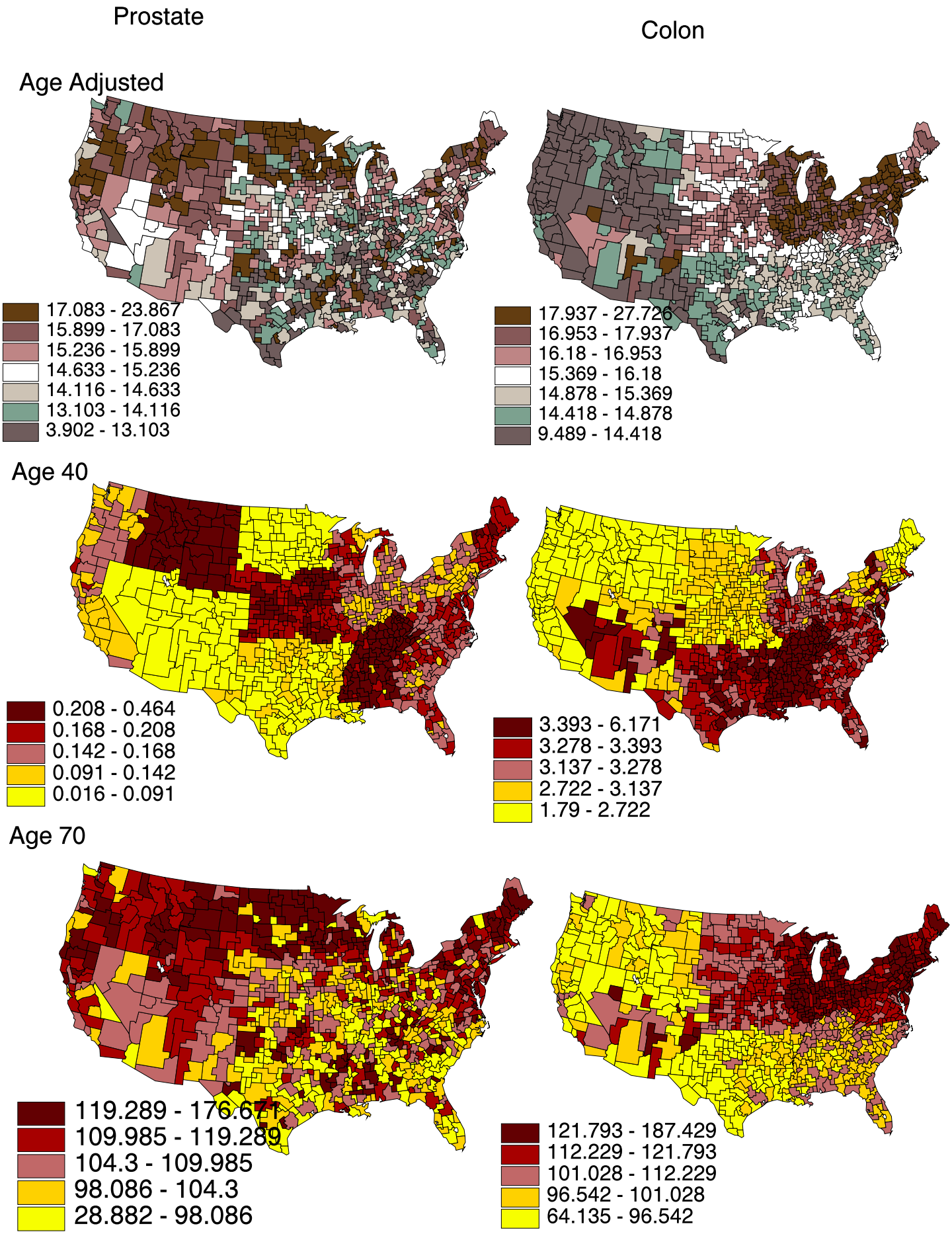
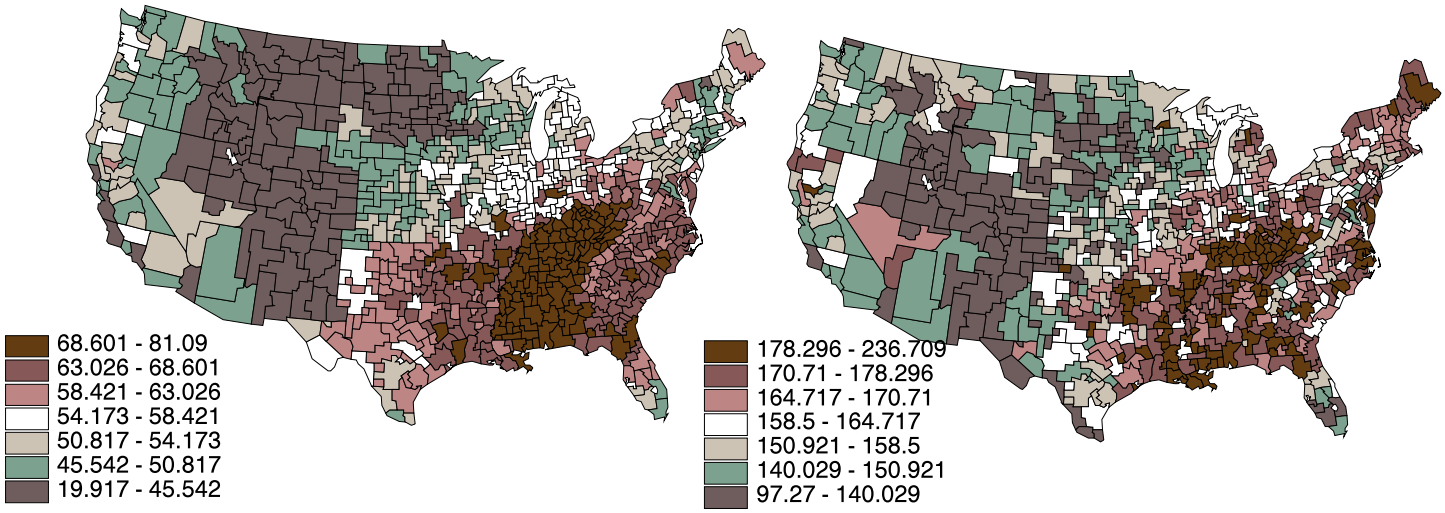


Fig. 3.8: Maps of the Estimated Death Rates by Type of Cancer

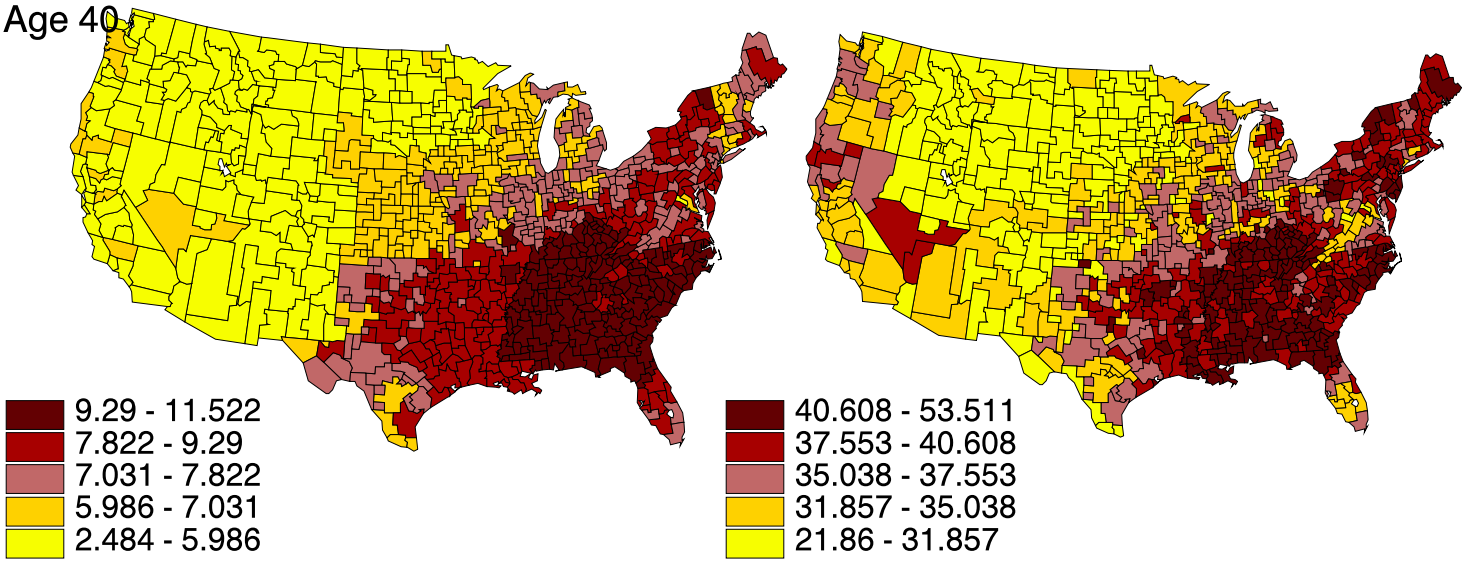
Lung

All Cancer

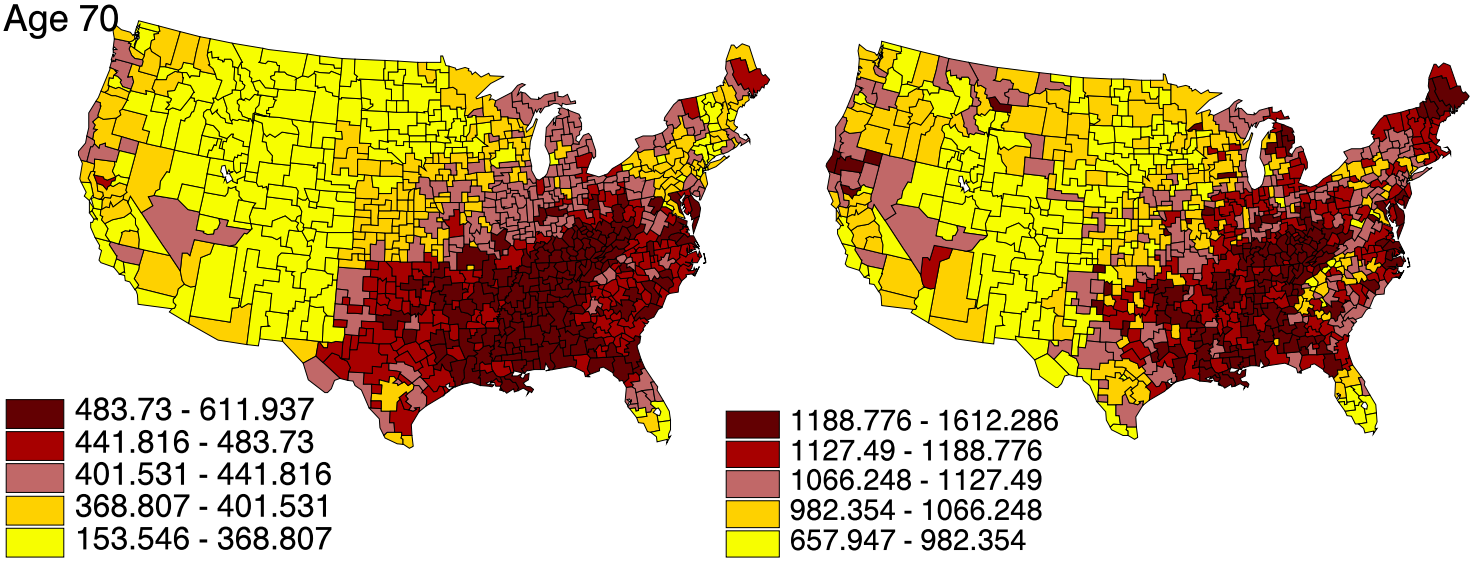
Age Adjusted



Age 40



Age 70



3.4 Analyses of Relative Occurences

3.4.1 Model Description

The model we fit is

$$\log\left(\frac{p_{ijk}}{p_{ij4}}\right) = \underline{Z_i'}\underline{\alpha} + \gamma_j + \eta_k + \theta_{jk} + \phi_i,$$

where $i = 1, \ldots, 798, \quad j = 1, \ldots, 10, \quad k = 1, 2, 3,$ and $\phi_i \sim N(0, \sigma_2^2)$, $\sigma_2^{-2} \sim \Gamma(\frac{a}{2}, \frac{b}{2})$ for $a = b = 0.002$.

Here $\underline{Z_i}$ is the matrix containing the 5 predictor variables such as per capita income, percentage of people below poverty level, education, epapm25 and epaso2. The variable income, epapm25 and epaso2 were divided by $10,000$ for computational stability. The $\hat{p}_{ijk}$ are the MLEs with an adjustment for zeros.

The corner point restrictions are $\eta_3 = 0$, $\gamma_1 = 0$, $\delta_{1k} = 0$, for $k = 1, 2, 3$, $\delta_{j3} = 0$, for $j = 2, \ldots, 7$.

We can define $\lambda_{ijk} = \lambda_{ij} p_{ijk}$ and since $\sum_{k=1}^{4} p_{ijk} = 1$ we have

$$p_{ijk} = \frac{\lambda_{ijk}}{\sum_{k=1}^{4} \lambda_{ijk}}.$$

Thus, we take

$$\hat{p}_{ijk}^{(t)} = \frac{\hat{\lambda}_{ijk}^{(t)}}{\sum_{k=1}^{4} \hat{\lambda}_{ijk}^{(t)}}$$

where $\hat{\lambda}_{ijk}^{(t)}$ for $t = 1, \ldots, 1000$ are the 1000 iterates from the Metropolis-Hastings samples.

Then, we fitted the model for each iteration $t = 1, \ldots, 1000$

$$\log\left(\frac{\hat{p}_{ijk}^{(t)}}{\hat{p}_{ij4}^{(t)}}\right) = \underline{Z_i'}\underline{\alpha}^{(t)} + \gamma_j^{(t)} + \eta_k^{(t)} + \theta_{jk}^{(t)} + \phi_i^{(t)}.$$

Once we obtained the least square estimates we then deduce $\phi_i^{(t)}$ such as

$$\phi_i^{(t)} = \frac{1}{3 * 10} \sum_{j,k} \left\{ \log\left(\frac{\hat{p}_{ijk}^{(t)}}{\hat{p}_{ij4}^{(t)}}\right) - \left(\underline{Z_i'}\underline{\alpha}^{(t)} + \gamma_j^{(t)} + \eta_k^{(t)} + \theta_{jk}^{(t)}\right) \right\}.$$

When estimates of $\underline{\alpha}, \underline{\gamma}, \underline{\eta}, \underline{\theta}$ are obtained, we use them to obtain $p_{ijk}^{(t)}$ again.

3.4.2 Estimates and Maps

Table 3.6 contains the mean, standard deviation and credible intervals of coefficients. Most of the covariates are significant and the estimates are different from the one obtained in Chapter 2 (see Table 2.2). The intercept and income are not significant anymore. Epapm25 and Epaso2 are the two most signifiacnt covariates. The parameters accounting for age class are all significant except for the second age class. The estimates of the interaction between age class and disease are clearly significant except for colon cancer at age class 2. The parameters accounting for type of cancer are significant.

In Figure 3.9 we present the distribution of the proportions of deaths p_{ijk} by disease. The patterns of the box plots across age classes look similar to the one obtained from the observed proportions in Figure 1.4.

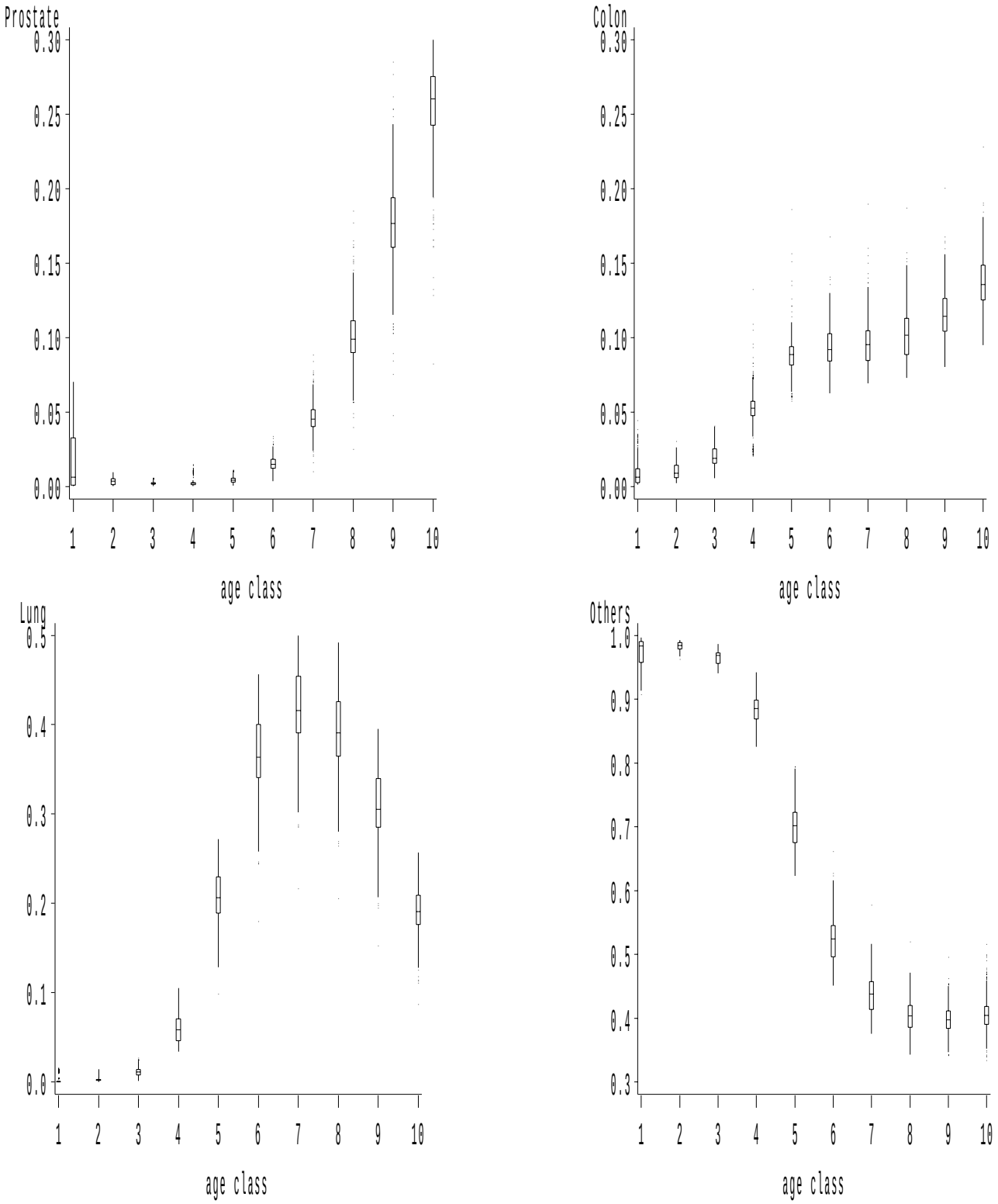


Figure 3.9: Distribution of the Proportions of Death by Disease

$\underline{\beta}$	Mean	Std	Interval
Intercept	0.013	0.008	[0.008, 0.018]
Income	-0.030	0.046	[-0.059, -0.000]
% Poverty	-0.002	0.001	[-0.003, -0.002]
College	0.001	0.000	[0.000, 0.001]
Epapm25	-6.571	2.171	[-8.052, -5.062]
Epaso2	2.610	0.463	[2.279, 2.920]
γ_2	0.001	0.000	[0.001, 0.001]
γ_3	0.009	0.001	[0.009, 0.009]
γ_4	0.065	0.002	[0.063, 0.066]
γ_5	0.297	0.005	[0.293, 0.300]
γ_6	0.710	0.006	[0.706, 0.714]
γ_7	0.974	0.006	[0.970, 0.978]
γ_8	0.986	0.005	[0.982, 0.989]
γ_9	0.780	0.004	[0.777, 0.783]
γ_{10}	0.471	0.005	[0.468, 0.474]
θ_{21}	-0.012	0.002	[-0.014, -0.011]
θ_{31}	-0.022	0.002	[-0.024, -0.020]
θ_{41}	-0.077	0.003	[-0.079, -0.075]
θ_{51}	-0.306	0.005	[-0.309, -0.302]
θ_{61}	-0.696	0.007	[-0.701, -0.691]
θ_{71}	-0.885	0.006	[-0.889, -0.881]
θ_{81}	-0.751	0.006	[-0.754, -0.747]
θ_{91}	-0.346	0.006	[-0.350, -0.342]
θ_{101}	0.180	0.009	[0.174, 0.185]
θ_{22}	0.001	0.001	[0.000, 0.002]
θ_{32}	0.004	0.001	[0.004, 0.005]
θ_{42}	-0.013	0.003	[-0.015, -0.011]
θ_{52}	-0.179	0.005	[-0.182, -0.176]
θ_{62}	-0.538	0.006	[-0.542, -0.534]
θ_{72}	-0.764	0.006	[-0.767, -0.760]
θ_{82}	-0.741	0.005	[-0.744, -0.738]
θ_{92}	-0.495	0.004	[-0.498, -0.493]
θ_{102}	-0.140	0.005	[-0.143, -0.136]
η_1	0.013	0.002	[0.011, 0.014]
η_2	0.006	0.001	[0.005, 0.006]

Table 3.6: Mean, Standard Deviation and 95% Credible Interval for the Estimates of the Parameters

Figures 3.10 and 3.11 show the empirical posterior densities for the 2 most significant covariates epapm25 and epaso2 and also for the variable accounting for the variation among HSAs, ϕ_i. They do not show much departure from the normal distribution.

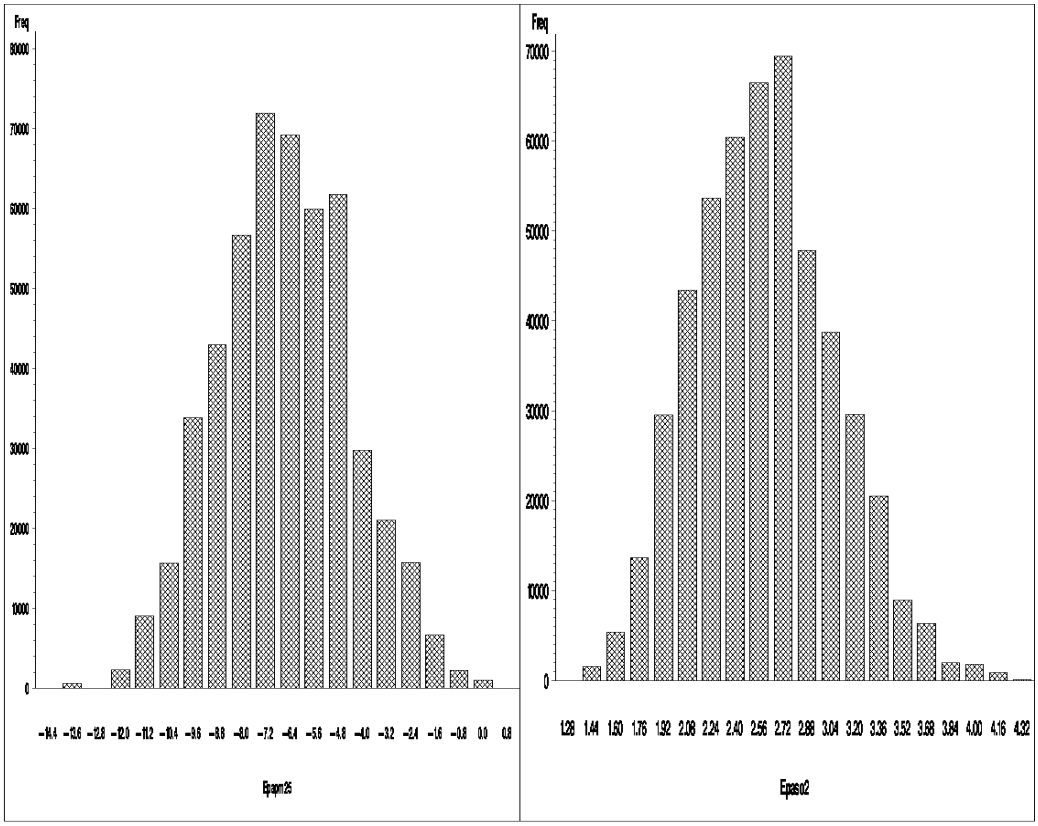


Figure 3.10: Frequency Histogram for Epapm25 and Epaso2

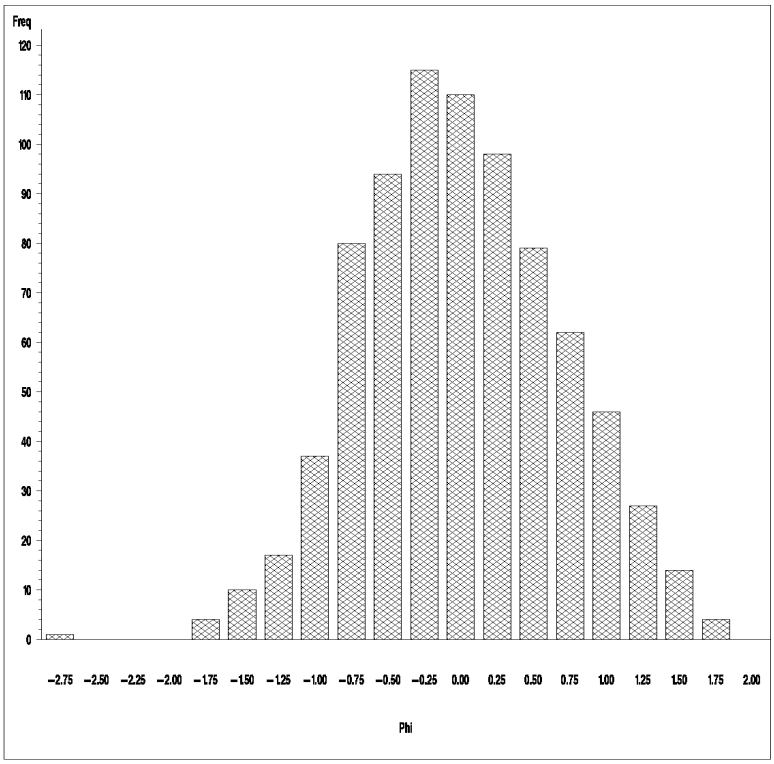


Figure 3.11: Frequency Histogram for Φ_i

In Figures 3.12, 3.13 and 3.14 we present the maps for age specific (40, 70 and 85 and up) of the proportions of deaths by disease. The regions of Pacific, Mountain North, Mountain South and West North Central-North present a concentration of high proportions of death for prostate cancer for people older than 65 years old and for other cancers for age 70 and younger. The Appalachian region and the South Atlantic-South present a concentration of high proportions of death for lung cancer for all age classes. The concentration of high proportions of death for colon cancer is in the North East and North Central East regions for age 70 and older.

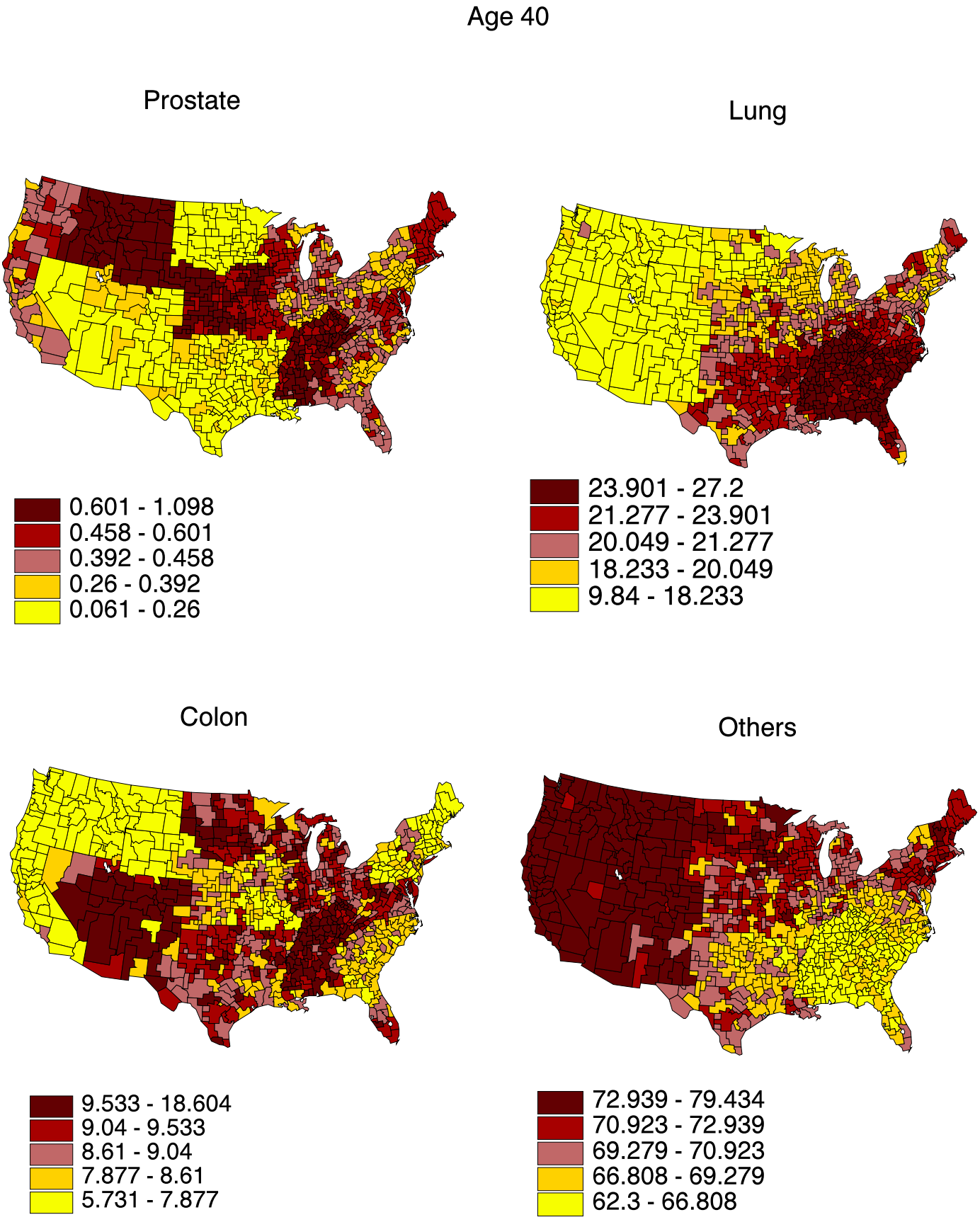


Fig. 3.12: Maps of the Proportions (10^-2) of Deaths for Age 40

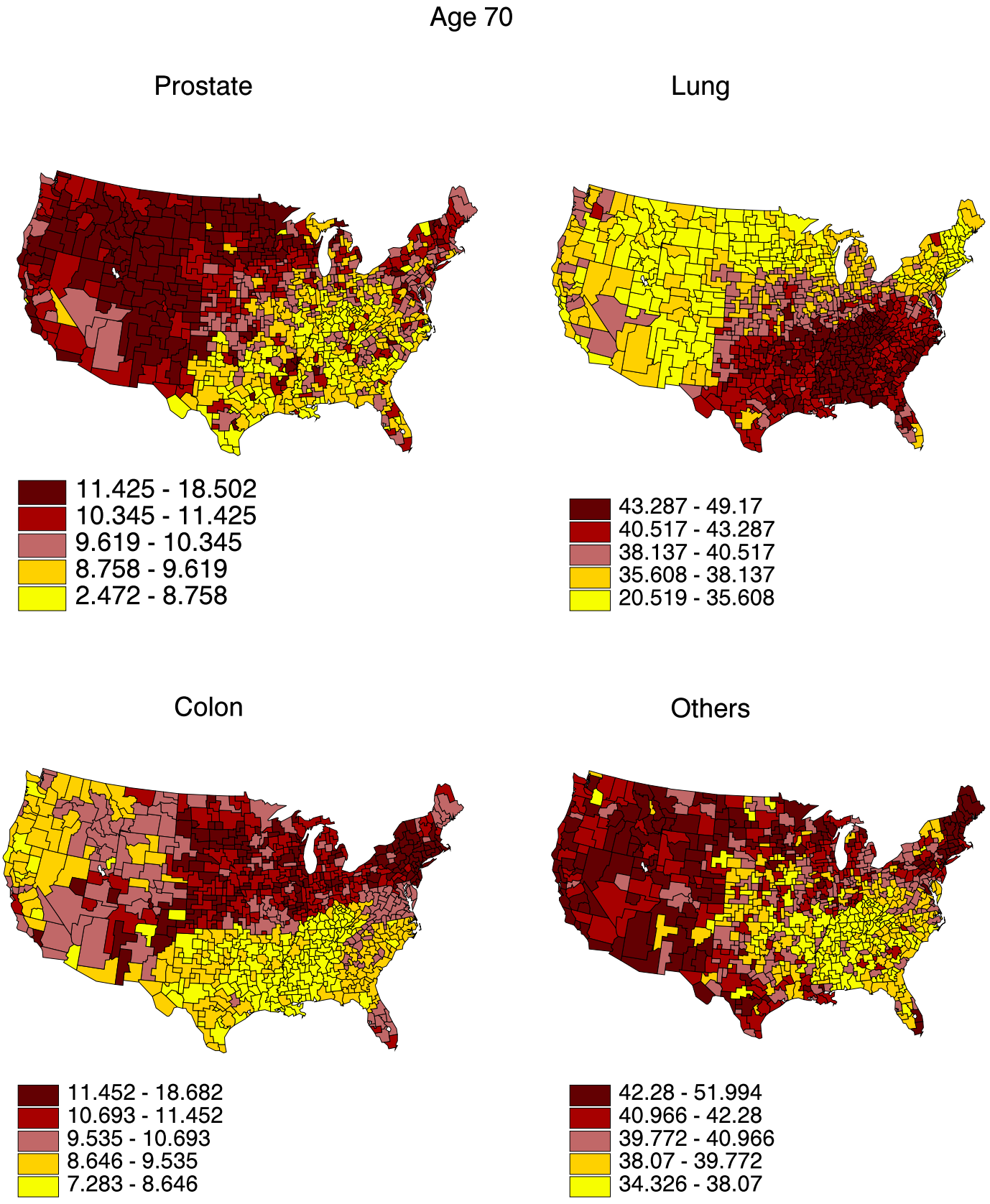


Fig. 3.13: Maps of the Proportions (10^-2) of Deaths for Age 70

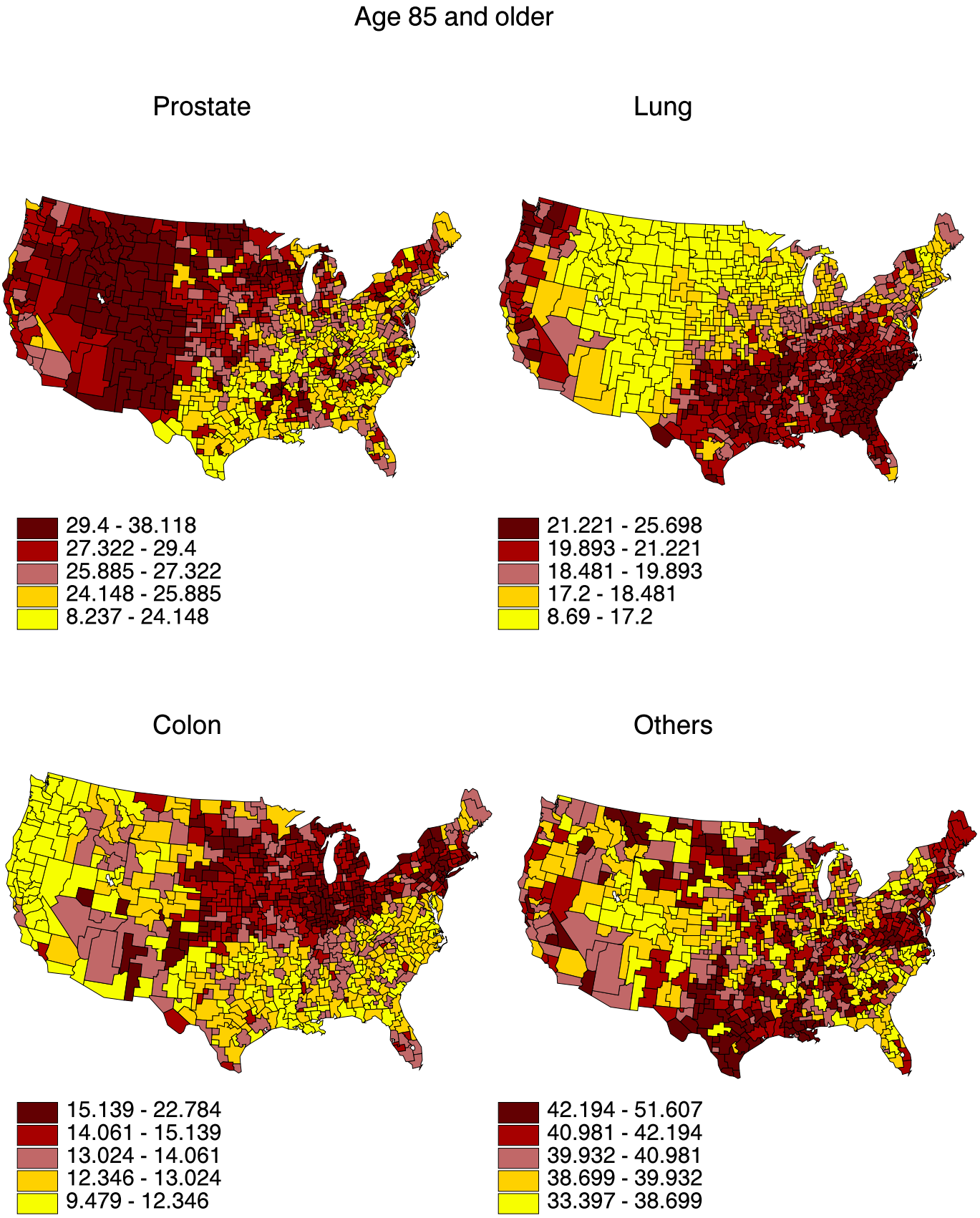


Fig. 3.14: Maps of the Proportions (10^-2) of Deaths for Age 85 and Above

The maps for age 40 for prostate and colon cancer are difficult to interpret since the data are very sparse. For prostate cancer, the regions of Mountain North, West North Central-South and East South Central concentrate the high proportions of death. For colon cancer, the high proportions of death are concentrated in the regions of Mountain South, West North Central-North and East South Central.

In Figures 3.15, 3.16, 3.17 and 3.18 we present the maps of the proportions of deaths for high and low values of epapm25 for prostate, colon, lung and other cancer respectively. In Figures 3.19, 3.20, 3.21 and 3.22 we present the corresponding maps for high and low values of epaso2. The maps do not show any strong pattern. Still, we observe that the proportion of deaths by lung cancer are high n the South Eastern states though there are high elsewhere as well.

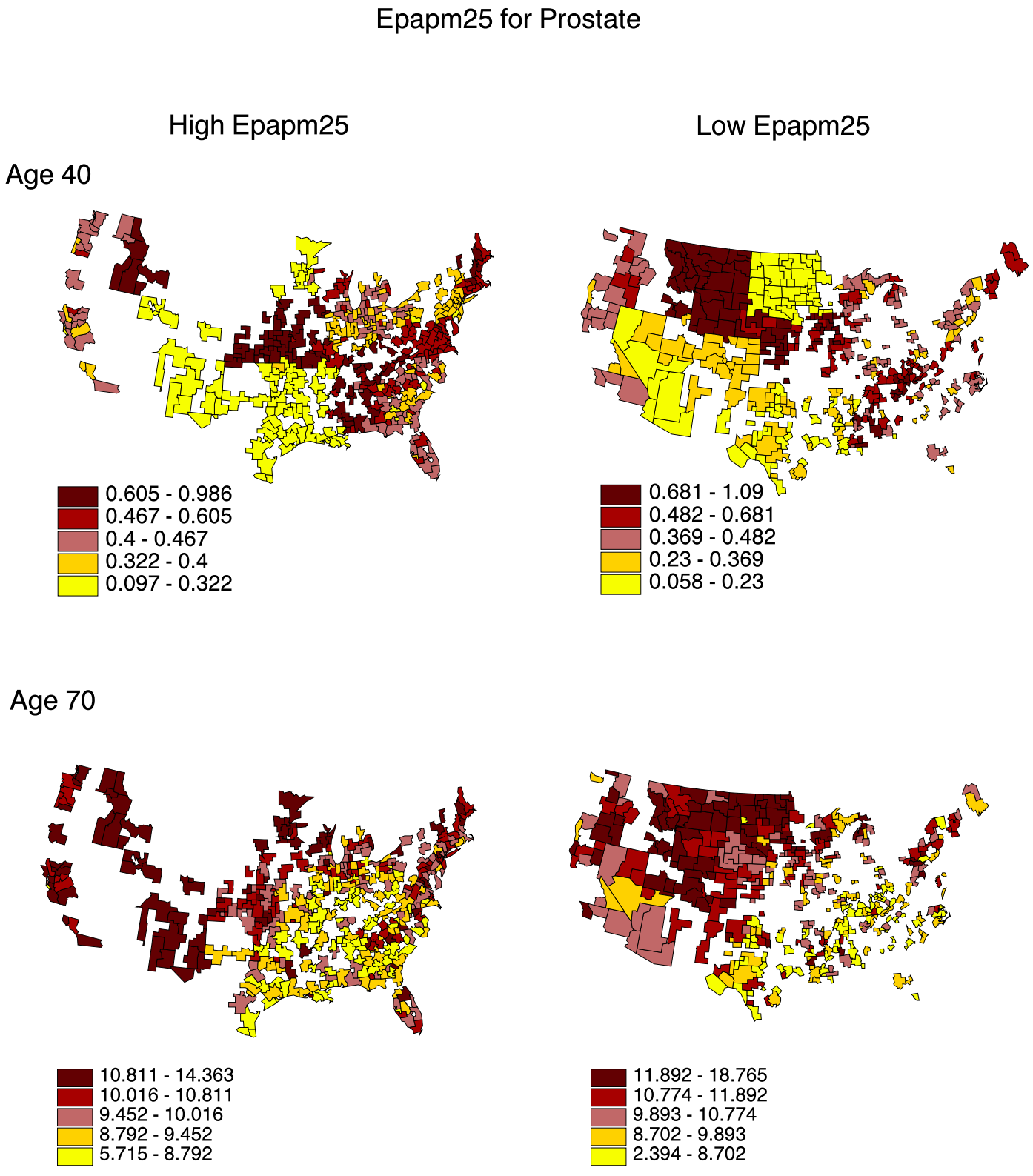


Fig. 3.15: Maps of the Proportions (10^-2) of Death for Prostate Cancer for High and Low Epapm25

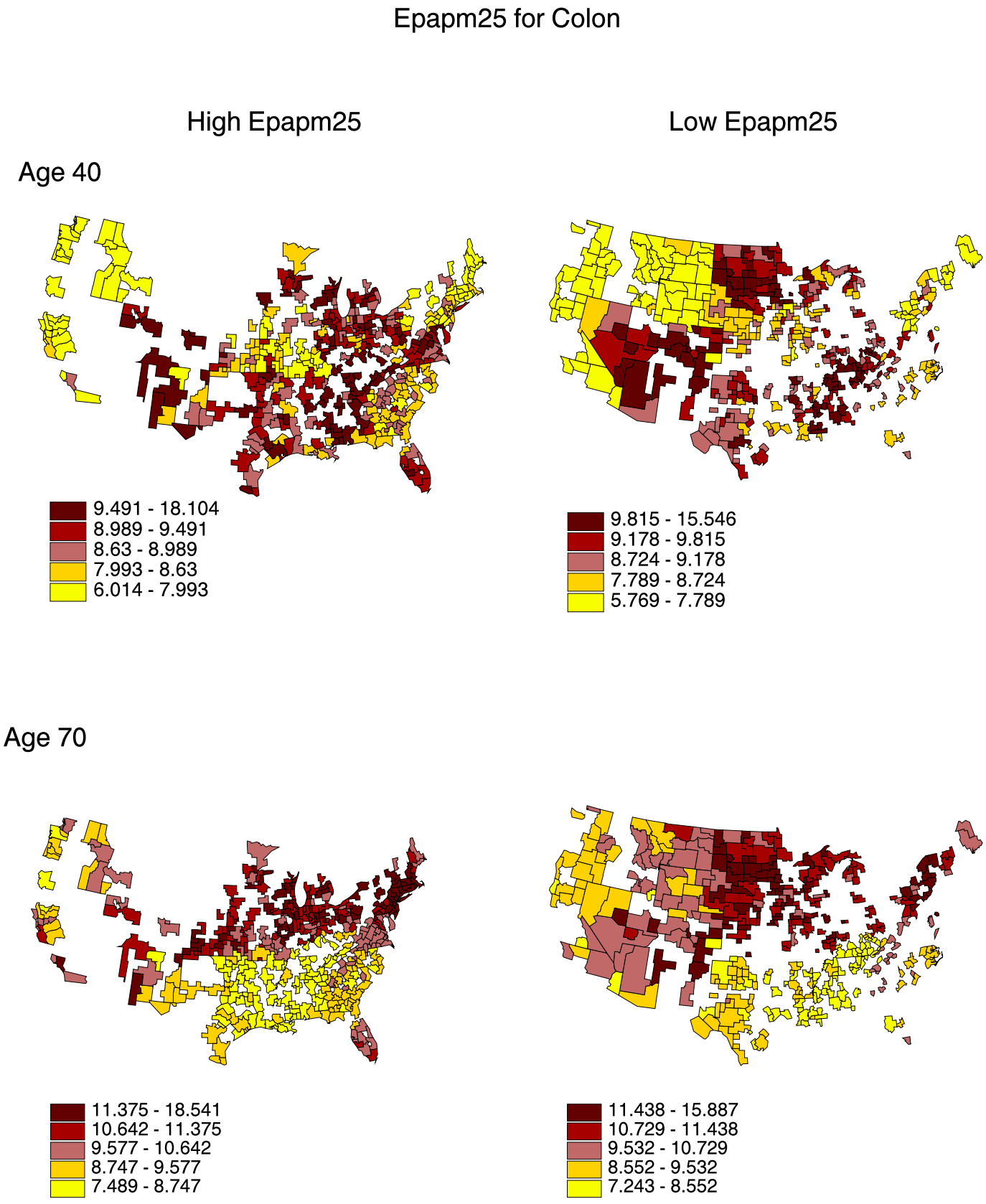


Fig. 3.16: Maps of the Proportions (10^-2) of Death for Colon Cancer for High and Low Epapm25

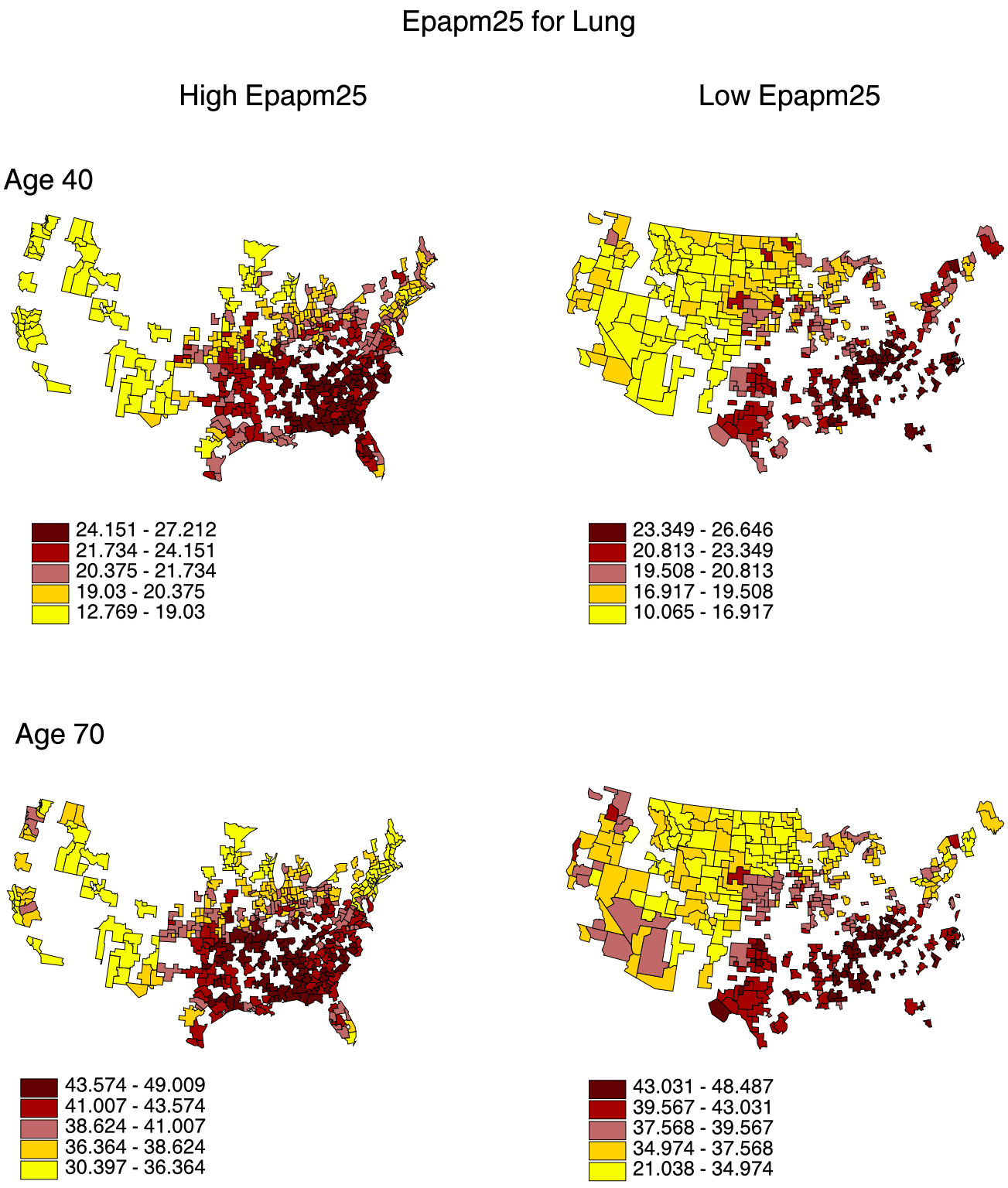


Fig. 3.17: Maps of the Proportions (10^-2) of Death for Lung Cancer for High and Low Epapm25

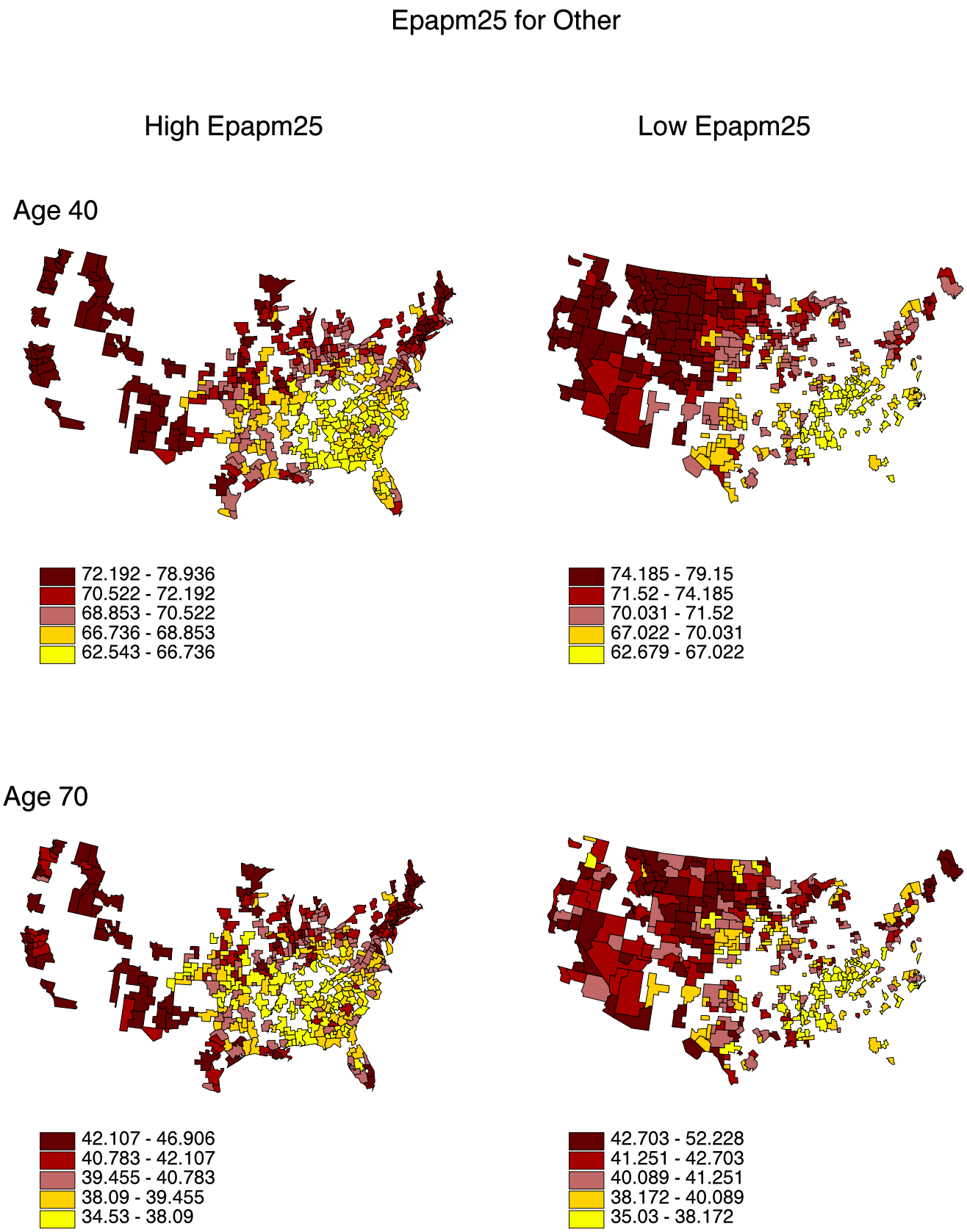


Fig. 3.18: Maps of the Proportions (10^-2) of Death for Other Cancer for High and Low Epapm25

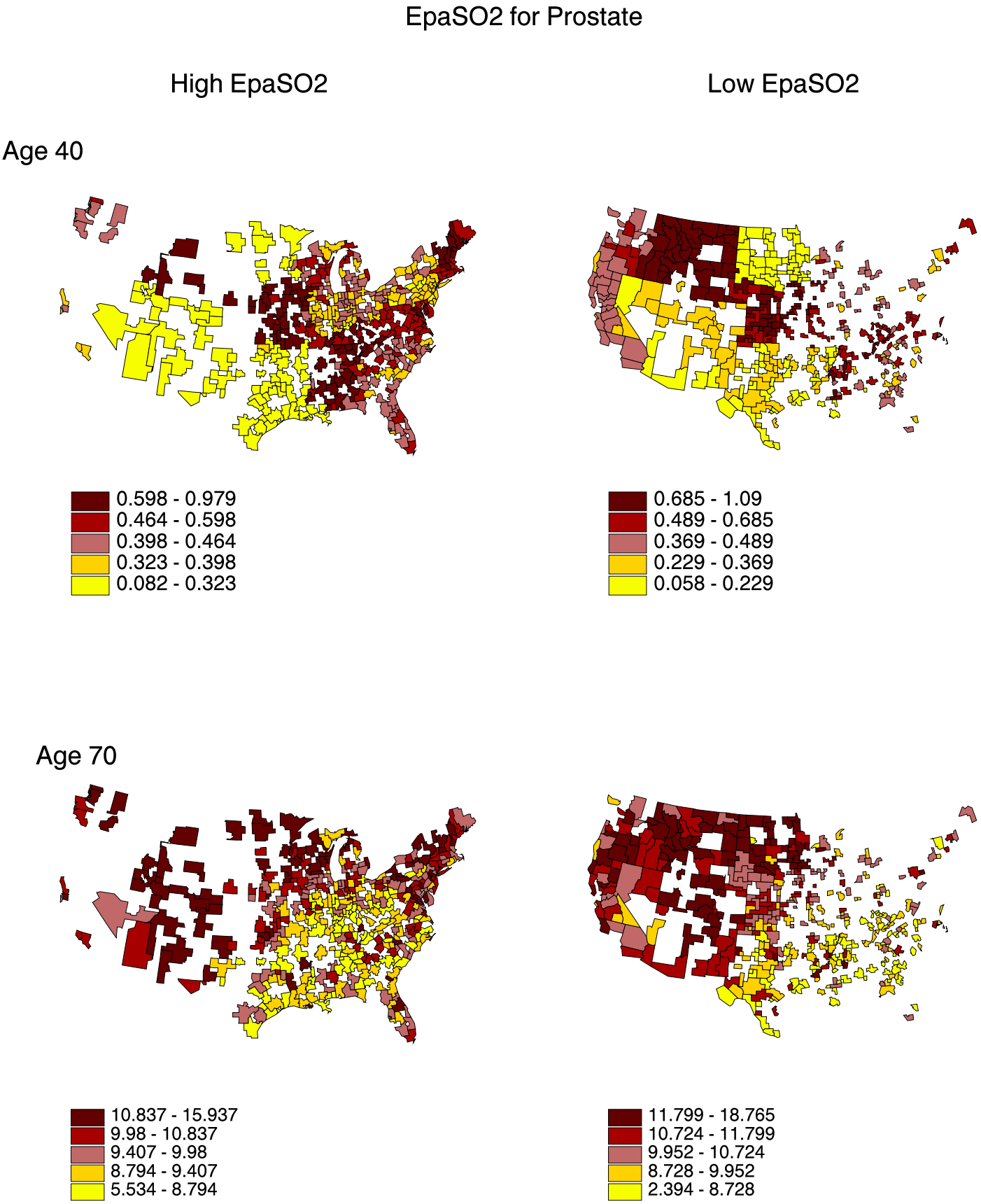


Fig. 3.19: Maps of the Proportions (10^-2) of Death for Prostate Cancer for High and Low EpaSO2

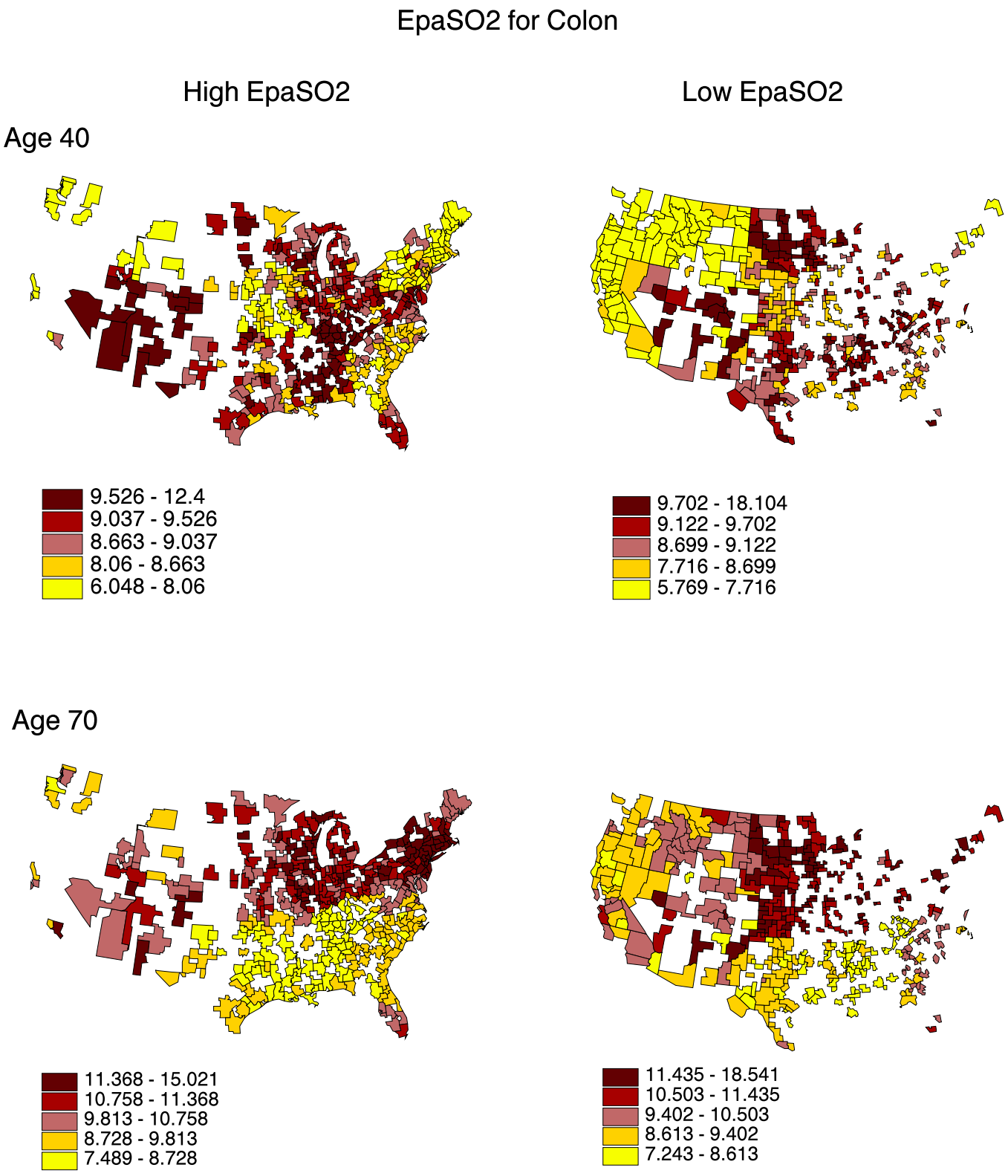


Fig. 3.20: Maps of the Proportions (10^-2) of Death for Colon Cancer for High and Low EpaSO2

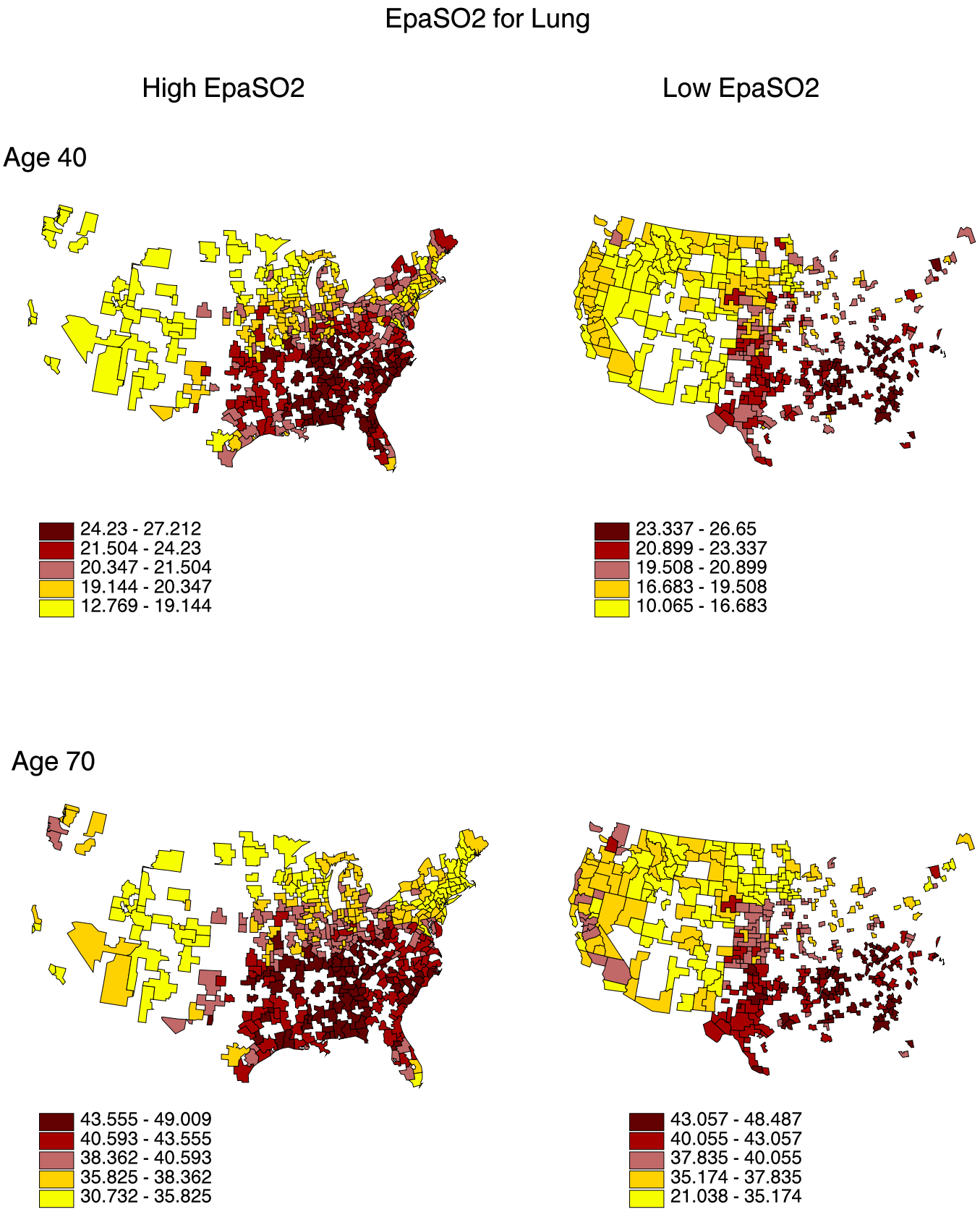


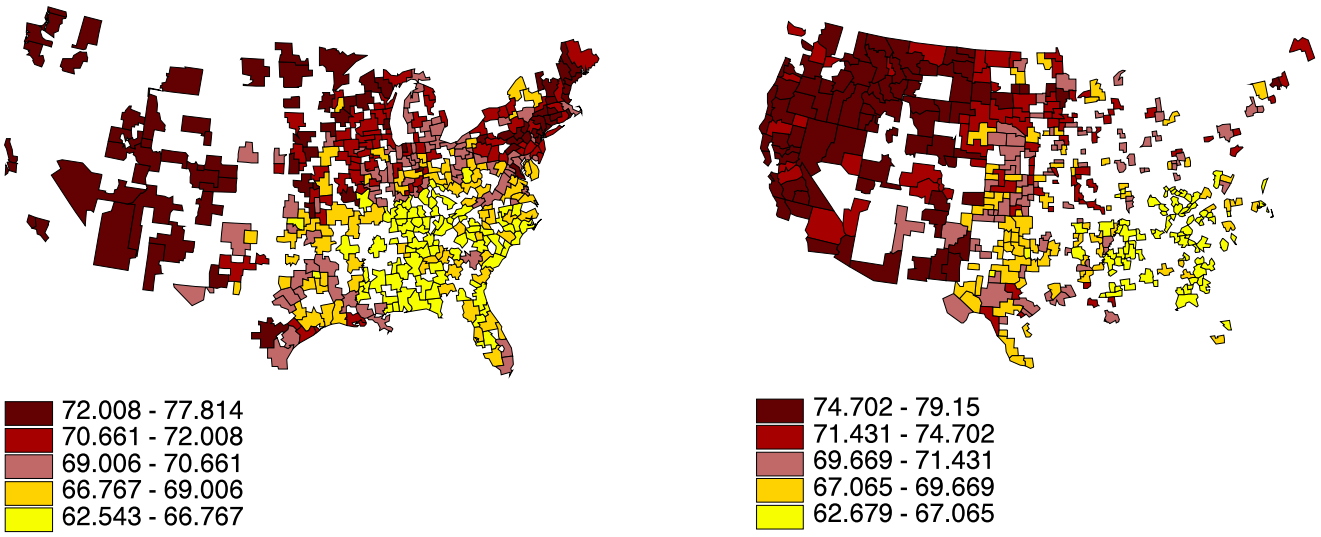
Fig. 3.21: Maps of the Proportions (10^-2) of Death for Lung Cancer for High and Low EpaSO2

EpaSO2 for Other

High EpaSO2

Low EpaSO2

Age 40



Age 70

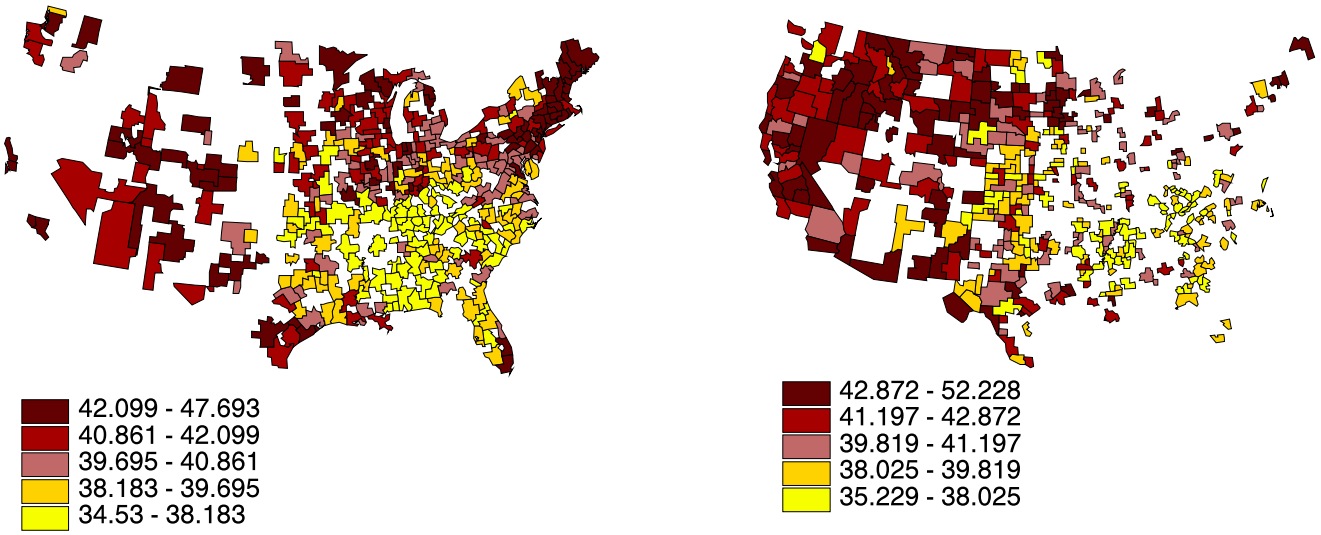


Fig. 3.22: Maps of the Proportions (10^-2) of Death for Other Cancer for High and Low EpaSO2

3.5 Concluding Remarks

First we fitted the NSP model to obtain the rates using the Metropolis-Hastings algorithm. Then we deduced the proportions of deaths for each disease and fitted a model which is a simple extension of the one discussed in Chapter 2 by incorporating a random effect for the HSAs.

The maps for different age classes and diseases show interesting patterns and highlighted some hot-spots. The maps for colon and prostate cancer for age less than 40 have to be interpreted with caution because of the sparseness of the data.

Chapter 4

Concluding Remarks

Our goals in this project are to analyze mortality rates for all types of cancer and to estimate their relative occurrences. We use Bayesian methods.

4.1 Review of Methodology

We make inference on the mortality rates by disease in two steps. Since our model shows that the mortality rates by disease are the product of the mortality rates for all cancer diseases and the proportions of death by disease, a two-part model is used.

To model the rates we used the NSP model of Nandram, Sedransk and Pickle (1999). Then we needed to develop new models for the occurrences.

Because the data are very sparse especially in the first 4 age classes, we amalgamated them as follows: age classes 1,2,3,4 as group 1, 5 as group 2, 6 as group 3, . . . , 10 as group 7.

We began by investigating the fit of an approximate model for the p_{ijk}. The parameter estimates, standard errors and maps were obtained. Unfortunately the maps had a pattern that was similar for each age class and each disease. The model

does not fit the data well. Then we fitted another approximate model. Only slight improvements were observed. The maps did not show any interesting pattern.

Finally, we apply the model used for the λ_{ij} to each disease so we got the estimates of the λ_{ijk} and then we deduce the p_{ijk}. Using Bayesian p-value and a cross-validation exercise, we showed that this model performs well. The maps for the mortality rates and for the relative occurrences showed different patterns for each disease across age classes. The maps obtained for age 40 should be interpreted with caution because of the sparseness of the data.

4.2 Final Results

The mortality rates for each type of cancer increase with age class, of course, with different intensities. Prostate cancer presents the widest range of mortality rates. The mortality rates for colon cancer increase across age classes but not as much as prostate and lung cancer.

The maps of the mortality rates show some interesting hot spots for prostate cancer in the North West and North Central regions. For colon cancer, the high mortality rates are concentrated in the regions of the North East and east North Central. The high mortality rates for lung cancer and all cancers are both concentrated in the South East region for each age class. This is not surprising since lung cancer is the leading cause of death by cancer.

The proportions of deaths for each type of cancer follow different patterns. The proportions of deaths increase with age class for prostate and colon cancer while it does so for lung cancer but only until age 55 where they begin to drop. Lung cancer

is the leading cause of cancer death across age class except for the older men (age 85 and above) for which the proportions are similar to the one for prostate cancer. One should note that the proportions of cancer deaths by prostate cancer increase steadily while they do so for colon cancer but slowly.

With all factors being the same, the odds of occurrences of prostate, colon and lung with respect to others decrease (increase) substantially with epapm25 (epaSO2). We conjecture that particulate matters in the air affects mostly young people but sulphur dioxide affects mostly the elderly.

The maps of the relative occurrences for age 40 for prostate and colon cancer are difficult to interpret since the data are very sparse. For prostate cancer, the regions of Mountain North, West North Central-South and East South Central concentrate the high proportions of death. For colon cancer, the high proportions of death are concentrated in the regions of Mountain South, West North Central-North and East South Central. The regions of Pacific, Mountain North, Mountain South and West North Central-North present a concentration of high proportions of deaths for prostate cancer for people older than 65 years old and for other cancers for age 70 and younger. The Appalachian region and the South Atlantic-South present a concentration of high proportions of deaths for lung cancer for all age classes. The concentration of high proportions of deaths for colon cancer is in the North East and North Central East regions for age 70 and older.

4.3 An Alternative Approach

We assume that $d_{ij}|n_{ij}, \lambda_{ij} \overset{ind}{\sim} Poisson(n_{ij}\lambda_{ij})$. We fit a hierarchical model with a single regression coefficient. The basis model for the analysis is as follows

$$\log \lambda_{ij} = \underline{x}_j^{'} \underline{\beta} + \nu_i + \delta_j \tag{4.1}$$

where $\underline{x}_j' = \big(1, decade_j, (decade_j)^2, (decade_j)^3, max\{0, (decade_j - knot)^3\}\big)$ with $decade_1 = 0.25$, $decade_j = j - 1$ for $j = 2, \ldots, 10$.

We assume that $\nu_i|\sigma_1^2 \overset{iid}{\sim} N\big(0, \sigma_1^2\big)$, $\delta_j|\sigma_2^2 \overset{iid}{\sim} N\big(0, \sigma_2^2\big)$ and the value of the knot that maximizes the likelihood of U.S. marginal data is 6 for "all cancer".

Here, $p(\underline{\beta}) = 1$ and $\sigma_{1k}^{-2}, \sigma_{2k}^{-2} \sim \Gamma(\frac{a}{2}, \frac{b}{2})$ where $a = b = 0.002$ to obtain a proper diffuse prior.

This can be fitted easily for all cancer. If one can model the individual type of cancer simultaneously through (4.1) the entire problem about the p_{ijk} would be solved automatically because

$$p_{ijk} = \frac{\lambda_{ijk}}{\lambda_{ij}}$$

such that the constraint $\lambda_{ij} = \sum_k \lambda_{ijk}$.

Better methods are needed to show variability in disease mapping over the areas. A model like the one in Chapter 2 but exact and with a random effect could incorporate better heterogeneity among areas. We attempted to do so already but with little success.

The real difficulty is in modeling the p_{ijk}. One could add a random effect in the model so we get the mixed effects model

$$\log\left(\frac{p_{ijk}}{p_{ij4}}\right) = \underline{Z}_i'\underline{\alpha} + \gamma_j + \eta_k + \theta_{jk} + \phi_i$$

where $i = 1, \ldots, 798$, $j = 1, \ldots, 10$, $k = 1, 2, 3$, $\underline{Z}_i$ containing the covariates, $\phi_i \sim N(0, \sigma_2^2)$, $\sigma_2^{-2} \sim \Gamma(.001, .001)$. The corner point restrictions are $\eta_3 = 0$, $\gamma_1 = 0$, $\theta_{1k} = 0$, for $k = 1, 2, 3$, $\theta_{j3} = 0$, for $j = 2, \ldots, 7$.

The Metropolis-Hastings algorithm was used but with little success. Since the estimates of the covariates contained in $\underline{\alpha}$ were not stable, we tried using one covariate

epapm25 and got better results. Besides that, the model was very sensitive to the sparseness of the data for the first four age classes so we tried to amalgamate the age classes so 5 age classes remain. The algorithm was working better but once we had got the estimates to move well, they were highly correlated.

It seems that the data do not permit this kind of model, but further investigation is required to fit this model.

Another aspect that would need to be addressed is to relate the variation among HSAs for the λ_{ij} and for the p_{ijk} by introducing a dependence such as

$$\begin{pmatrix} \nu_i \\ \phi_i \end{pmatrix} \overset{iid}{\sim} N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Delta \right).$$

Yet another important point would be to incorporate some spatial structure on $\begin{pmatrix} \nu_i \\ \phi_i \end{pmatrix}$. In this case one would need to delete the intercept from both models.

Bibliography

[1] Chib, S. and Greenberg, E. (1995). "Understanding the Metropolis-Hastings Algorithm." *The American Statistician*, 49, 327-335.

[2] Gelfand, A.E., Sahu, S.K., and Carlin, B.P. (195). "Efficient Reparametrizations For Normal Linear Models," *Biometrika*, 82, 479-488.

[3] Gelman, A., Carlin,J.B., Stern, H.S., Rubin D.B. (1995). *Bayesian Data Analysis*. London: Chapman and Hall.

[4] Gideon, A. (1999). "Bayesian Analysis and Mapping of Breast Cancer Mortality Data for U.S. Health Service Areas," M.S. Thesis, Mathematical Sciences, Worcester Polytechnic Institute, Worcester, Massachusetts.

[5] Hettmansperger, T. (1984). *Statistical Inference Based on Ranks*. New York: Wiley.

[6] Nandram, B., Sedransk, J., and Pickle, L.(1999). "Bayesian Analysis of Mortality Rates for U.S. Health Service Areas," *Sankhya*, Vol. 61, 145-165.

[7] Nandram, B., Sedransk, J., and Pickle, L.(1999a). "Bayesian Analysis and Mapping of Mortality Rates for Chronic Obstructive Pulmonary Disease," under review.

[8] Pickle, L., Mungiole, M., Jones, G.K., and White, R. Crouchley (1996). *Atlas of U.S. Mortality*. Hyattsville, MD: National Center for Health Statistics.

[9] Pickle, L., Mungiole, M., Jones, G.K., and White A.A. (1997). "Analysis of Mapped Mortality Data by Mixed Effect Models," Technical Report, National Center for Health Statistics.

Appendix A: Model for the λ_{ij}

A locally uniform prior distribution is used on $\underline{\beta}$ for each region and a proper diffuse prior on σ_1^2 and σ_2^2 as follows

$$p(\underline{\beta}) = 1 \text{ and } \sigma_1^2, \sigma_2^2 \sim \Gamma(\frac{a}{2}, \frac{b}{2}) \text{ where } a = b = 0.002.$$

To fit this model, we use the Metropolis-Hastings algorithm implementing the product of Kernels Principle (Chib and Greenberg 1995) which essentially allows us to draw successively from each conditional posterior distribution, instead of having to run each of the conditional posterior distribution to convergence for every value of the conditional variables (parameters). Whenever an opportunity arises, we also use the technique of centering (e.g. Gelfand et al. 1995) to facilitate computations. An independent chain is used in the Metropolis-Hastings step which permits blocking.

The key idea to obtain the proposal density in any of our Metropolis steps is to use a second order Taylor's series expansion about a convenient point (an approximation to the mode) for each conditional posterior distribution. Letting $M = 10$ and $N = 798$, the joint posterior density is

$$\begin{aligned} p(\underline{\beta}, \underline{\nu}, \underline{\delta}, \sigma_1^2, \sigma_2^2 | \underline{d}) \quad &\propto \quad \prod_{i=1}^{N} \prod_{j=1}^{M} \left\{ e^{(\underline{x}_j' \underline{\beta} + \nu_i + \delta_j)\, d_{ij} \,-\, n_{ij}\, e^{\underline{x}_j' \underline{\beta} + \nu_i + \delta_j}} \right\} \\ &\times \quad \prod_{i=1}^{N} \left\{ \left(\frac{1}{\sigma_1^2} \right)^{1/2} e^{-\frac{1}{2\sigma_1^2} \nu_i^2} \right\} \prod_{j=1}^{M} \left\{ \left(\frac{1}{\sigma_2^2} \right)^{1/2} e^{-\frac{1}{2\sigma_2^2} \delta_j^2} \right\} \\ &\times \quad \left(\frac{1}{\sigma_1^2} \right)^{a/2+1} e^{-\frac{b}{2\sigma_1^2}} \left(\frac{1}{\sigma_2^2} \right)^{a/2+1} e^{-\frac{b}{2\sigma_2^2}}. \end{aligned}$$

First of all, we make the transformation $\underline{x}_j'\underline{\beta} + \delta_j = \phi_j$ which leads to

$$\begin{aligned} p(\underline{\beta}, \underline{\nu}, \underline{\delta}, \sigma_1^2, \sigma_2^2|\underline{d}) \;\; &\propto \;\; \prod_{i=1}^{N}\prod_{j=1}^{M}\left\{ e^{(\nu_i+\phi_j)\, d_{ij} \,-\, n_{ij}\, e^{\nu_i+\phi_j}} \right\} \\ &\times \;\; \prod_{i=1}^{N}\left\{ \left(\frac{1}{\sigma_1^2}\right)^{1/2} e^{-\frac{1}{2\sigma_1^2}\nu_i^2} \right\} \prod_{j=1}^{M}\left\{ \left(\frac{1}{\sigma_2^2}\right)^{1/2} e^{-\frac{1}{2\sigma_2^2}(\phi_j - \underline{x}_j'\underline{\beta})^2} \right\} \\ &\times \;\; \left(\frac{1}{\sigma_1^2}\right)^{a/2+1} e^{-\frac{b}{2\sigma_1^2}} \left(\frac{1}{\sigma_2^2}\right)^{a/2+1} e^{-\frac{b}{2\sigma_2^2}}. \end{aligned}$$

Then, we can deduce

$$\sigma_1^{-2}|\underline{\beta}, \underline{\nu}, \underline{\phi}, \sigma_2^2, \underline{d} \;\sim\; \Gamma\left(\frac{N+a}{2}, \frac{b+\sum_i \nu_i^2}{2}\right) \tag{2}$$

$$\sigma_2^{-2}|\underline{\beta}, \underline{\nu}, \underline{\phi}, \sigma_1^2, \underline{d} \;\sim\; \Gamma\left(\frac{M+a}{2}, \frac{b+\sum_j (\phi_j - \underline{x}_j'\underline{\beta})^2}{2}\right) \tag{3}$$

$$\underline{\beta}|\underline{\nu}, \underline{\phi}, \sigma_1^2, \sigma_2^2, \underline{d} \;\sim\; N\left(\Big(\sum_j \underset{\sim j}{x}\, \underline{x}_j'\Big)^{-1} \Big(\sum_j \phi_j \underset{\sim j}{x}\Big), \sigma_2^2 \Big(\sum_j \underset{\sim j}{x}\, \underline{x}_j'\Big)^{-1} \right) \tag{4}$$

$$p(\phi_j|\underline{\beta}, \underline{\nu}, \sigma_1^2, \sigma_2^2, \underline{d}) \;\sim\; \prod_i \left\{ e^{(\nu_i+\phi_j)\, d_{ij} \,-\, n_{ij}\, e^{\nu_i+\phi_j}} \right\} e^{-\frac{1}{2\sigma_2^2}(\phi_j - \underline{x}_j'\underline{\beta})^2} \tag{5}$$

$$p(\nu_i|\underline{\beta}, \underline{\phi}, \sigma_1^2, \sigma_2^2, \underline{d}) \;\sim\; \prod_j \left\{ e^{(\nu_i+\phi_j)\, d_{ij} \,-\, n_{ij}\, e^{\nu_i+\phi_j}} \right\} e^{-\frac{1}{2\sigma_1^2}\nu_i^2}. \tag{6}$$

Since the conditional posterior densities (5) and (6) are difficult to work with, we use the Metropolis algorithm in these two cases.

First, we consider conditional posterior density of $\phi_j|\underline{\beta}, \underline{\nu}, \sigma_1^2, \sigma_2^2, \underline{d}$. We denote $\Delta(\phi_j)$ the logarithm of the conditional posterior densities of $\phi_j|\underline{\beta}, \underline{\nu}, \sigma_1^2, \sigma_2^2, \underline{d}$ such as

$$\begin{aligned} \Delta(\phi_j) \;\; &= \;\; \sum_i \left\{ (\nu_i + \phi_j)\, d_{ij} \;-\; n_{ij}\, e^{\nu_i+\phi_j} \right\} - \frac{1}{2\sigma_2^2}(\phi_j - \underline{x}_j'\underline{\beta})^2 \\ &= \;\; A(\phi_j) - \frac{1}{2\sigma_2^2}(\phi_j - \underline{x}_j'\underline{\beta})^2. \end{aligned}$$

Then,

$$\begin{aligned} \frac{dA(\phi_j)}{d\phi_j} \;\; &= \;\; \sum_i \left(d_{ij} \;-\; n_{ij}\, e^{\nu_i+\phi_j} \right) \\ \frac{d^2A(\phi_j)}{d\phi_j^2} \;\; &= \;\; -\sum_i n_{ij}\, e^{\nu_i+\phi_j}. \end{aligned}$$

Thus, based on $A(\phi_j)$ alone we can get an estimator of ϕ_j by setting $\frac{dA(\phi_j)}{d\phi_j} = 0$:

$$\hat{\phi_j} = \log\left\{\frac{\sum_i d_{ij}}{\sum_i n_{ij} e^{\nu_i}}\right\}.$$

For the Metropolis step, we take

$$\phi_j|\underline{\beta}, \underline{\nu}, \sigma_1^2, \sigma_2^2, \underline{d} \overset{approx}{\sim} Uniform(a, b),$$

where

$$\begin{cases} a = E\big(\phi_j|\underline{\beta}, \underline{\nu}, \sigma_1^2, \sigma_2^2, \underline{d}\big) - k * SD\big(\phi_j|\underline{\beta}, \underline{\nu}, \sigma_1^2, \sigma_2^2, \underline{d}\big) \\ b = E\big(\phi_j|\underline{\beta}, \underline{\nu}, \sigma_1^2, \sigma_2^2, \underline{d}\big) + k * SD\big(\phi_j|\underline{\beta}, \underline{\nu}, \sigma_1^2, \sigma_2^2, \underline{d}\big). \end{cases}$$

Second, we consider how to draw $\underline{\nu}$. We denote the logarithm of the conditional posterior distribution of $\nu_i|\underline{\beta}, \underline{\phi}, \sigma_1^2, \sigma_2^2, \underline{d}$ by $\Delta(\nu_i)$ where

$$\begin{aligned} \Delta(\nu_i) &= \sum_j \left\{(\nu_i + \phi_j)\, d_{ij} - n_{ij}\, e^{\nu_i + \phi_j}\right\} - \frac{1}{2\sigma_1^2}\nu_i \\ &= A(\phi_j) - \frac{1}{2\sigma_1^2}\nu_i. \end{aligned}$$

Then,

$$\begin{aligned} \frac{dA(\nu_i)}{d\nu_i} &= \sum_j \left(d_{ij} - n_{ij}\, e^{\nu_i + \phi_j}\right) \\ \frac{d^2A(\nu_i)}{d\nu_i^2} &= -\sum_j n_{ij}\, e^{\nu_i + \phi_j}. \end{aligned}$$

Thus, based on $A(\nu_i)$ alone we can get an estimator of ν_i by setting $\frac{dA(\nu_i)}{d\nu_i} = 0$:

$$\hat{\nu_i} = \log\left\{\frac{\sum_j d_{ij}}{\sum_j n_{ij} e^{\nu_i}}\right\}$$

and we can deduce that

$$\nu_i|\underline{\beta}, \underline{\phi}, \sigma_1^2, \sigma_2^2, \underline{d} \overset{approx}{\sim} Uniform(a, b),$$

where

$$\begin{cases} a = E\big(\nu_i|\underline{\beta}, \underline{\phi}, \sigma_1^2, \sigma_2^2, \underline{d}\big) - k * SD\big(\nu_i|\underline{\beta}, \underline{\phi}, \sigma_1^2, \sigma_2^2, \underline{d}\big) \\ b = E\big(\nu_i|\underline{\beta}, \underline{\phi}, \sigma_1^2, \sigma_2^2, \underline{d}\big) + k * SD\big(\nu_i|\underline{\beta}, \underline{\phi}, \sigma_1^2, \sigma_2^2, \underline{d}\big). \end{cases}$$

APPENDIX B: **First Approximate Model for the p_{ijk}**

Matrix Formulation of the Model

Let's denote $\underline{y}_i = \begin{pmatrix} y_{i,1,1} \\ y_{i,1,2} \\ y_{i,1,3} \\ \vdots \\ y_{i,j,1} \\ y_{i,j,2} \\ y_{i,j,3} \\ \vdots \\ y_{i,7,1} \\ y_{i,7,2} \\ y_{i,7,3} \end{pmatrix}$, $x_i = \begin{pmatrix} \underline{Z}_i' & \\ \vdots & A \\ \underline{Z}_i' & \end{pmatrix}$

where $A = \left(\begin{array}{cccc|cccccc|cccc|cc} 0 & 0 & \dots & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 & \dots & 0 & 0 & 1 & 0 \\ 0 & 0 & \dots & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 1 \\ 0 & 0 & \dots & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 \\ 1 & 0 & \dots & 0 & 1 & 0 & \dots & 0 & 0 & 0 & 0 & \dots & 0 & 0 & 1 & 0 \\ 1 & 0 & \dots & 0 & 0 & 0 & \dots & 0 & 0 & 1 & 0 & \dots & 0 & 0 & 0 & 1 \\ 1 & 0 & \dots & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 \\ 0 & 1 & \dots & 0 & 0 & 1 & \dots & 0 & 0 & 0 & 0 & \dots & 0 & 0 & 1 & 0 \\ 0 & 1 & \dots & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 1 & \dots & 0 & 0 & 0 & 1 \\ 0 & 1 & \dots & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & 1 & 0 & 0 & \dots & 1 & 0 & 0 & 0 & \dots & 0 & 0 & 1 & 0 \\ 0 & 0 & \dots & 1 & 0 & 0 & \dots & 0 & 0 & 0 & 0 & \dots & 0 & 1 & 0 & 1 \\ 0 & 0 & \dots & 1 & 0 & 0 & \dots & 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 \end{array}\right)$.

Then, the model can be written as

$$\underline{Y} = X\underline{\beta} + \underline{\epsilon} \quad \text{where } \underline{\epsilon} \sim N\big(\underline{0}, \sigma^2 W\big) \text{ and } \underline{\beta} = \Big(\underset{6*1}{\underline{\alpha}}\ \gamma_2\ \gamma_3\ \dots\ \gamma_7\ \delta_{21}\ \dots\ \delta_{72}\ \eta_1\ \eta_2\Big)'. \tag{7}$$

The weight matrix W is a block diagonal matrix of (21×21) matrices W_i which are themselves block diagonal matrices of (3×3) matrices w_{ij}.

The estimators are such as

$$\hat{\underline{\beta}} = \big(X'W^{-1}X\big)^{-1}X'W^{-1}Y,$$

$$Cov\big(\hat{\underline{\beta}}\big) = \big(X'W^{-1}X\big)^{-1}\hat{\sigma}^2,$$

$$\hat{\sigma}^2 = \left(\underline{Y} - X\underline{\hat{\beta}}\right)'\left(\underline{Y} - X\underline{\hat{\beta}}\right) \Big/ (21 \times n - 26).$$

For computational purposes, we did not use matrix algebra to invert the matrices but the following identities and transformations

$$X'W^{-1}X = \sum_{i=1}^{n} X_i' W_i^{-1} X_i \text{ and } X'W^{-1}Y = \sum_{i=1}^{n} X_i' W_i^{-1} Y_i$$

.

Construction of the Weights

Since we assume that the d_{ijk}'s follow a multinomial distribution such as

$$\underline{d}_{ij}|\underline{p}_{ij} \overset{ind}{\sim} Multinomial(d_{ij.}, \underline{p}_{ij}) \text{ and } \hat{p}_{ijk} = \frac{d_{ijk}}{d_{ij.}} \implies \frac{\hat{p}_{ijk}}{\hat{p}_{ij3}} = \frac{d_{ijk}}{d_{ij3}},$$

$$\text{then Var} \begin{pmatrix} \hat{p}_{ij1} \\ \hat{p}_{ij2} \\ \hat{p}_{ij3} \\ \hat{p}_{ij4} \end{pmatrix} = \frac{1}{d_{ij.}} \begin{bmatrix} p_{ij1}(1-p_{ij1}) & -p_{ij1}p_{ij2} & -p_{ij1}p_{ij3} & -p_{ij1}p_{ij4} \\ -p_{ij1}p_{ij2} & p_{ij2}(1-p_{ij2}) & -p_{ij2}p_{ij3} & -p_{ij2}p_{ij4} \\ -p_{ij1}p_{ij3} & -p_{ij2}p_{ij3} & p_{ij3}(1-p_{ij3}) & -p_{ij3}p_{ij4} \\ -p_{ij1}p_{ij4} & -p_{ij2}p_{ij4} & -p_{ij3}p_{ij4} & p_{ij4}(1-p_{ij4}) \end{bmatrix}$$

Now, let $\underline{T}$ be a vector with finite variance Σ and mean $\underline{\mu}$. Take any function $f(\underline{T})$, then the first order Taylor's series expansion of $f(\underline{T})$ is

$$f(\underline{T}) \approx f(\underline{\mu}) + (\underline{T} - \underline{\mu})'\underline{G}(\underline{\mu}) \text{ where } \underline{G}(\underline{\mu}) \text{ is the gradient vector.}$$

Therefore,

$$\begin{aligned} \text{Var}\left(f(\underline{T})\right) &\approx \text{Var}\left((\underline{T} - \underline{\mu})'\underline{G}(\underline{\mu})\right) \\ &= \underline{G}(\underline{\mu})^T \text{ Var}(\underline{T})\ \underline{G}(\underline{\mu}), \end{aligned}$$

and

$$\text{Cov}\left(f_1(\underline{T}), f_2(\underline{T})\right) \approx \underline{G_{f_1}}(\underline{\mu})^T \text{ Var}(\underline{T})\ \underline{G_{f_2}}(\underline{\mu}).$$

Our three functions are $f_k(\underline{p}) = \log p_{ijk} - \log p_{ij4}$ for $k = 1, \ldots, 3$.

Then $\mathrm{Var}\left(f_k(\hat{\underline{p}})\right) = \frac{1}{d_{ij.}}\left(\frac{1}{p_{ijk}} + \frac{1}{p_{ij4}}\right) \approx \frac{1}{d_{ijk}} + \frac{1}{d_{ij4}}$, and

$$\mathrm{Cov}\left(f_1(\underline{p}), f_2(\underline{p})\right) = \mathrm{Cov}\left(f_1(\underline{p}), f_3(\underline{p})\right) = \mathrm{Cov}\left(f_2(\underline{p}), f_3(\underline{p})\right) = \frac{1}{d_{ij.}p_{ij4}} \approx \frac{1}{d_{ij4}}.$$

It follows that $w_{ij} = \begin{pmatrix} \frac{1}{d_{ij1}} + \frac{1}{d_{ij4}} & \frac{1}{d_{ij4}} & \frac{1}{d_{ij4}} \\ \frac{1}{d_{ij4}} & \frac{1}{d_{ij2}} + \frac{1}{d_{ij4}} & \frac{1}{d_{ij4}} \\ \frac{1}{d_{ij4}} & \frac{1}{d_{ij4}} & \frac{1}{d_{ij3}} + \frac{1}{d_{ij4}} \end{pmatrix}$.

Approximations

Approximations on the responses and the weights

In order to compute $\log\left(\frac{p_{ijk}}{p_{ij4}}\right)_{p_{ijk}=\hat{p}_{ijk}} = \log\left(\frac{d_{ijk}}{d_{ij4}}\right)$, $i = 1, 2, 3$, we used the new definition d^*_{ijk} of d_{ijk}

$$d^*_{ijk} = \begin{cases} 10^{-6} & \text{if } d_{ijk} = 0, \\ d_{ijk} & \text{otherwise.} \end{cases}$$

Moreover, for small d_{ijk}, Pickle et al. (1996) have shown that one will obtain better estimates of the λ_{ij} and the p_{ijk} using more stable quantities obtained which can be obtained computing the averages by regions for small d_{ijk}'s such as

$$d^{**}_{ijk} = \begin{cases} \frac{\sum_{i\in R}\sum_{j=1}^{10} d_{ijk}}{\sum_{i\in R}\sum_{j=1}^{10}\sum_k d_{ijk}} \frac{\sum_{i\in R}\sum_k d_{ijk}}{n_r} & \text{if } d_{ijk} < 3, \\ d_{ijk} & \text{if } d_{ijk} \geq 3, \end{cases}$$

where $R =$ region and $n_r =$ number of HSAs in the region R.

The last substitution is motivated by

$$\frac{\sum_{i\in R}\sum_{j=1}^{10} d_{ijk}}{\sum_{i\in R}\sum_{j=1}^{10}\sum_k d_{ijk}} \approx \frac{d_{ijk}}{\sum_k d_{ijk}} \quad \text{and } \frac{\sum_{i\in R}\sum_k d_{ijk}}{n_r} \approx \sum_k d_{ijk} \ .$$

Methodology

After fitting model (7), we can deduce the p_{ijk}'s by writing the model as follows

$$\begin{bmatrix} log(\frac{p_{ij1}}{p_{ij4}}) \\ log(\frac{p_{ij2}}{p_{ij4}}) \\ log(\frac{p_{ij3}}{p_{ij4}}) \\ \vdots \end{bmatrix} = X\underline{\beta} = \begin{bmatrix} \theta_{ij1} \\ \theta_{ij2} \\ \theta_{ij3} \\ \vdots \end{bmatrix},$$

where

$$\begin{cases} p_{ij1} & = e^{\theta_{ij1}}/(1+e^{\theta_{ij1}}+e^{\theta_{ij2}}+e^{\theta_{ij3}}) \\ p_{ij2} & = e^{\theta_{ij2}}/(1+e^{\theta_{ij1}}+e^{\theta_{ij2}}+e^{\theta_{ij3}}) \\ p_{ij3} & = e^{\theta_{ij3}}/(1+e^{\theta_{ij1}}+e^{\theta_{ij2}}+e^{\theta_{ij3}}) \\ p_{ij4} & = 1/(1+e^{\theta_{ij1}}+e^{\theta_{ij2}}+e^{\theta_{ij3}}). \end{cases} \tag{8}$$

Then, we have obtained the estimates $\hat{\underline{\beta}}$ and $\widehat{Cov(\hat{\underline{\beta}})} = \hat{\Sigma}$ and we have approximated the distribution of $\hat{\underline{\beta}}$ by

$$\hat{\underline{\beta}} \ \sim \ N_{26}\Big(\underline{\beta}, \hat{\Sigma}\Big).$$

It is convenient to use a Bayesian approach. Then by taking the non-informative prior for $\underline{\beta}$, i.e $p(\underline{\beta}) = 1$, it is obvious that approximately

$$\underline{\beta}|\underline{d} \ \sim \ N_{26}\Big(\hat{\underline{\beta}}, \hat{\Sigma}\Big). \tag{9}$$

Therefore we can draw a sample of $1,000$ $\underline{\beta}$'s from (9). For computational purposes, we partitioned the parameters as follows

$$\underset{26*1}{\underline{\beta}} = \begin{bmatrix} \underset{10*1}{\underline{\beta}_1} \\ \hline \underset{8*1}{\underline{\beta}_2} \\ \hline \underset{8*1}{\underline{\beta}_3} \end{bmatrix} \quad \text{and} \quad \underset{26*26}{\hat{\Sigma}} = \left[\begin{array}{cc|c} \underset{10*10}{\hat{\Sigma}_{11}} & \underset{10*8}{\hat{\Sigma}_{12}} & \underset{10*8}{\hat{\Sigma}_{13}} \\ \underset{8*10}{\hat{\Sigma}_{21}} & \underset{8*8}{\hat{\Sigma}_{22}} & \underset{8*8}{\hat{\Sigma}_{23}} \\ \hline \underset{8*10}{\hat{\Sigma}_{31}} & \underset{8*8}{\hat{\Sigma}_{32}} & \underset{8*8}{\hat{\Sigma}_{33}} \end{array}\right]. \tag{10}$$

Then, we

i) generate $\underline{\beta}_1$ using the marginal distribution $\underline{\beta}_1|\underline{b}_1 \ \sim \ N_8\Big(\hat{\underline{\beta}}_1, \hat{\Sigma_{11}}\Big)$,

ii) generate the conditional distribution of $\underline{\beta}_2|\underline{\beta}_1 = \underline{b}_1$ which is normal and has the mean

$$\hat{\underline{\beta}}_2 + \hat{\Sigma_{21}}\hat{\Sigma_{11}}^{-1}(\underline{b}_1 - \hat{\underline{\beta}}_1),$$

and covariance

$$\hat{\Sigma_{22}} - \hat{\Sigma_{21}}\hat{\Sigma_{11}}^{-1}\hat{\Sigma_{12}}$$

,

iii) and finally generate the conditional distribution of $\underline{\beta_3}|\underline{\beta_1} = \underline{b_1}, \underline{\beta_2} = \underline{b_2}$ which is Normal and has the mean

$$\underline{\hat{\beta}_3} + (\hat{\Sigma_{31}}, \hat{\Sigma_{32}}) \begin{pmatrix} \hat{\Sigma_{11}} & \hat{\Sigma_{12}} \\ \hat{\Sigma_{21}} & \hat{\Sigma_{22}} \end{pmatrix}^{-1} \begin{pmatrix} \underline{b_1} - \underline{\hat{\beta}_1} \\ \underline{b_2} - \underline{\hat{\beta}_2} \end{pmatrix},$$

and covariance

$$\hat{\Sigma_{33}} - (\hat{\Sigma_{31}}, \hat{\Sigma_{32}}) \begin{pmatrix} \hat{\Sigma_{11}} & \hat{\Sigma_{12}} \\ \hat{\Sigma_{21}} & \hat{\Sigma_{22}} \end{pmatrix}^{-1} \begin{pmatrix} \hat{\Sigma_{13}} \\ \hat{\Sigma_{23}} \end{pmatrix}$$

.