Supplemental Materials

Socioeconomic data

Thirteen variables from the China Statistical Yearbook (1986 and 1997) are selected to represent each of the five key factor groups - income, education, employment, medical resources, and pollution which could influence the LCM rate ([1](#_ENREF_1),[2](#_ENREF_2)). Province level smoking prevalence rate was extracted from two smoking surveys conducted by the Committee of National Patriotic Health Campaign (CNPHC) and Ministry of Health (MOH), in 1984 and 1996 respectively. All variables are converted to percent or per capita values (based on population data of the year) for comparisons among provincial level regions of different sizes. At the county level, we use census population data to further study the relationship between the main factors selected by the .provincial level study. After 1949, China conducted six population censuses in 1953, 1964, 1982, 1990, 2000 and 2010. Detailed population data before the last LCM survey in 2004-05 were reported by the 1982, 1990, and 2000 censuses. The 1964 census reported a limited number of measures on the general population, of which the percent of non-agriculture population was used as an indicator of economic development level in China in the 1960s. The 1953 census was not used because it only reported on a few general indicators of population ([3](#_ENREF_3)).

Units of analysis: province or county

Based on data availability, both province and county level analyses are conducted in this study. China changed provincial boundaries in northern part of the country by distributing parts of Inner Mongolia among surrounding provinces in 1969. The provincial boundaries changed back to the previous provincial boundaries in 1979 which are also the boundaries today. This change only affected 1973-75 mortality survey’s provincial boundaries but we did not do provincial level analysis for the time period due to no data for the regional factors during and before the time period. The 1990-92 and 2004-05 mortality surveys and the population censuses were all done during the time that the provincial boundaries are the same as today. Other changes of provincial boundaries are Hainan province that was separated from Guangdong province in 1988 and Chongqing municipality that was separated from Sichuan province in 1997. Since our earliest provincial level regional factor data was in 1984 that was before the Hainan province separated from Guangdong, we merged the mortality data of Guangdong and Hainan together in 1990-92 and 2004-05, and Chongqing and Sichuan together in 2004-05. The current province boundary is shown in Supplemental Figure 1.

Although county level administrative regions are quite stable, some changes of the county boundaries or names did occur during the time period. Due to the change of administrative boundaries for counties and cities as well as the change of reporting units from an entire city as one unit (e.g. Beijing as one unit) in the 1982 and 1990 population censuses to each urban district as one unit (e.g. Dongcheng District in Beijing as one unit) in the 2000 population census, we have to exclude 22 locations (several cities had multiple urban districts surveyed) from the 1990-92 mortality survey in order to match the 1982 and 1990 population census data for our analyses. There are no multiple urban districts in the same city in the 2004-05 mortality survey dataset, so all 158 locations are used for the analyses.

All county level locations in four population censuses and three mortality survey datasets are geo-coded using the county level GB (Guo Biao or National Standard) Codes (GB2260-91) for administrative divisions in 1:1M China administrative regions county level (1 July 1990) GIS map produced by the Socioeconomic Data and Applications Center of the Center for International Earth Science Information Network (CIESIN).([4](#_ENREF_4)) We used 1990’s GIS map instead of the most recent one to accommodate the county boundaries and names during the entire time period of our study because 1990 is about in the middle of the study period. We then spatially verified every location to ensure that all the mortality and census population data for the same geographic unit are coded with the same GB code, particularly for the regions that had name, administration level or boundary changes.

Poisson regression model and cluster analysis at the county level:

The observed number of deaths, (for county *i* and sex *j*), are viewed as a realization from the random variable,, which has the conditional Poisson distribution *Poisson*() with the county population , and the overall rate *R* for the whole country. The product is the expected number of deaths for the county, assuming the whole country has the same rate, *R*. Parameter can be interpreted as the risk ratio (RR) of the county’s observed deaths over its expected deaths. Poisson regression uses the log-function to link the effect of any covariate to the death by



where represents a vector of county level risk factors, such as SES, smoking, healthcare availability and accessibility, environmental pollution, etc. The parameter is then interpreted as the logarithm of RR for the factor of interest. In other words,  is interpreted as the RR. ([5](#_ENREF_5)) The Poisson regression model is flexible in that random effects can be included in addition to the fixed effect model, so that dependence on the space and/or time scale can be considered. This forms the basis of the generalized linear mixed effect models. In our analysis, we used the SAS PROC GLIMMIX procedure ([6](#_ENREF_6)) to include the county level spatial autocorrelation in the exponential form. The exponential form of the spatial random effect is a kind of distance-based autocorrelation function. Like any other distance-based functions, it specifies a stronger spatial correlation between regions closer to each other, i.e., the spatial correlation coefficient decreases exponentially with the distance between two regions. As the distance exceeds a certain value, the correlation coefficient is assumed to disappear. ([6](#_ENREF_6)) Other forms of distance-based functions, such as Gaussian function and spherical function, are based on the same principle and results from them are usually very similar to those from the exponential form. Calculation of other forms involves more parameters, making the algorithm hard to converge. Hence exponential form is usually selected as the autocorrelation function. Another choice, adjacent neighbor matrix, assumes that neighboring regions (sharing boundary lines) have stronger spatial autocorrelation. It is suitable for regions with more regular shape, size, and length of boundary lines. Sizes, shapes, and boundary lines between counties or provinces in China are highly variable, so adjacency is not an optimal choice compared to distance-based autocorrelation functions.

Area level risk factors are selected using forward selection with a retaining criterion of p-value=0.2. Pseudo R2 is calculated to measure model fit. ([7](#_ENREF_7)) Itis calculated as the coefficient of determination in the log-transformation scale between the death counts and the model predicted death counts.

The spatial scan statistics assumes conditional independent Poisson random counts, the same way as in the Poisson regression model. When scanning for clusters *without* model adjustment, it is assumed that death count is proportional to the county population size, i.e., the RR is not dependent on risk factors and is fixed at 1 across all regions. The input population data is the county population in the SaTScan setting in this case. For cluster analysis *with* model adjustment, the contributing risk factors in the Poisson regression model predict the death count. In the SaTScan setting, instead of county population, the county’s Poisson model predicted death count is set as the population input, so any identified clusters will be detected above what is explained in the Poisson regression, and point to further analysis of additional risk factors. Elliptical and circular clusters are searched to allow for different shapes of clusters. The maximum spatial window size is set at the default of 50% population, and clusters are reported using the GINI index criterion ([8](#_ENREF_8)) for the 1973-75 data. The hierarchical most likely clusters are reported for the 1990-92 and 2004-05 data.

GINI criterion ([8](#_ENREF_8)) is a cluster reporting criterion implemented since SaTScan version 9.2. Along with p-value, cluster location, size, RR, which are reported in the default SaTScan report, it also reports the GINI index value. Literature search ([8](#_ENREF_8)) found that most research papers applying SaTScan lack criterion in selecting the spatial scan window, an important parameter in deciding the cluster locations and sizes. As a result, many choose the default of 50% of population, which unavoidably reports unnecessary and meaningless large clusters with small relative risks. The GINI criterion in SaTScan borrows the Lorenz curve and GINI index concept from the economics field to measure the disparity in health disparity (LCM in this study), so any reported clusters are selected based on the maximum GINI value and hence the optimal differentiation between clusters and non-clusters.

The cluster analysis (both without and with model adjustment) is based on the county-level Poisson regression model whose results are presented in Table 3. Depend variables are death counts by sex, county, and survey, and independent variables are those reported as significant with p-value <.05 (presented with point estimate and 95% CI for RR in Table 3). The independent variables are selected in a stepwise selection process and those with p-value<.05 are selected to be included in the model for the cluster analysis. The selected model does affect the detected clusters, so it is important that the model does not leave out key covariates, meanwhile not too many covariates are present to avoid the problem of collinearity over-parametrization. In this study, the stepwise variable selection is performed, in which every step adds one covariate and eliminate another possible candidate, until no further addition or elimination of covariates is possible according to a pre-defined criteria (p-value<.05). This procedure results in a robust set of detected clusters, which is not affected by selected models given the available data.

Supplemental Table 1 Variables and data source for each of the analysis in this study

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variable | Measure  (abbreviation if any) | | Year available | Geographic aggregation level | Data source |
| Dependent | Lung cancer death count  by gender | | 1973-75 | County | mortality survey, China CDC, MOH |
| 1990-1992 |
| 2004-2005 |
| 2013 | Province | Global Health Data Exchange |
| Population  and County-level Covariates | Male and female population (pop) | | All censuses | County and Province | Population censuses |
| % Non-Agricultural Population in Total Pop (% non-agr pop) | | 1964 |
| % of Senior High School and Above (HS edu) | | 1982 |
| % Urban Population (Urban) | |
| % of Industry Labor in Total Pop (% Ind Labor) | | 1982 and 1990 |
| Province-level  Covariates  by  Category | Income | GDP per capita (GDP p.c.) | 1984-85  and  1997 | Province | China Statistical Yearbook |
| Consumption per Capita (Cons. p.c.) |
| Employment | % of Agriculture Labor in Total Pop (% Agr. Labor) |
| % of Industry Labor in Total Pop (% Ind. Labor) |
| % of Construction Labor in Total Pop （% Cons. Labor） |
| % of Service Labor in Total Pop （% Serv. Labor） |
| Education | Literacy Rate (>15 years old) (Literacy) |
| % of Senior High School and Above (HS edu) |
| Medical Resources | Hospital Beds per 1000 Persons (Hsptl bed) |
| Medical Personnel per 1000 Persons (Med. Persn.) |
| Pollution | Waste Water per Capita (Waste Water p.c.) |
| Gas Emission per Capita  (Gas p.c.) |
| Solid Waste per Capita (Solid Waste p.c.) |
| Smoking | % of Total Smoker in Total Pop (% Total Smoker) | 1984 and 1996 | Province | Smoking Survey conducted by CNPHC and MOH |
| % of Male Smoker in Total Pop (% Male Smoker) |
| % of Female Smoker in Total Pop (% Female Smoker) |

Supplemental Table 2. Correlations Between LCM Rates (1990-2013) and Area Factors (1984-1996), province level

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1984-85 | 1990-92 Lung Cancer Death Rate | | | 2004-05 Lung Cancer Death Rate | | | 2013 Lung Cancer Death Rate | | |
| Area Factors | 26 regions | | | 29 regions | | | 29 regions | | |
|  | Total | Men | Women | Total | Men | Women | Total | Men | Women |
| GDP p. c. | 0.788\*\* | 0.794\*\* | 0.711\*\* | 0.616\*\* | 0.645\*\* | 0.525\*\* | 0.169 | 0.103 | 0.291 |
| Cons. p.c. | 0.807\*\* | 0.765\*\* | 0.809\*\* | 0.614\*\* | 0.582\*\* | 0.62\*\* | 0.167 | 0.067 | 0.362\* |
| % Agri. Labor | -0.737\*\* | -0.666\*\* | -0.793\*\* | -0.691\*\* | -0.634\*\* | -0.724\*\* | -0.282 | -0.171 | -0.497\*\* |
| % Ind. Labor | 0.822\*\* | 0.805\*\* | 0.782\*\* | 0.741\*\* | 0.754\* | 0.658\*\* | 0.311\* | 0.235 | 0.438\*\* |
| % Cons. Labor | 0.561\*\* | 0.477\*\* | 0.66\*\* | 0.524\*\* | 0.467\*\* | 0.575\*\* | 0.134 | 0.053 | 0.294 |
| % Serv. Labor | 0.607\*\* | 0.543\*\* | 0.669\*\* | 0.512\*\* | 0.469\*\* | 0.544\*\* | 0.037 | -0.048 | 0.216 |
| Literacy | 0.547\*\* | 0.485\*\* | 0.604\*\* | 0.685\*\* | 0.655\*\* | 0.658\*\* | 0.566\*\* | 0.511\*\* | 0.614\*\* |
| HS edu | 0.692\*\* | 0.653\*\* | 0.701\*\* | 0.636\*\* | 0.619\*\* | 0.613\*\* | 0.195 | 0.115 | 0.343\* |
| Hsptl Bed | 0.752\*\* | 0.722\*\* | 0.73\*\* | 0.349\* | 0.316\* | 0.382\* | 0.053 | -0.038 | 0.253 |
| Med. Persn | 0.062 | 0.086 | 0.012 | 0.406\* | 0.421\* | 0.33\* | 0.725\*\* | 0.761\*\* | 0.573\*\* |
| Waste Water p.c. | 0.686\*\* | 0.694\*\* | 0.617\*\* | 0.656\*\* | 0.691\*\* | 0.541\*\* | 0.325\* | 0.268 | 0.358\* |
| Gas p.c. | 0.444\* | 0.352\* | 0.568\*\* | 0.433\* | 0.347\* | 0.528\*\* | 0.143 | 0.064 | 0.266 |
| Solid Waste p.c. | 0.44\* | 0.368\* | 0.515\*\* | 0.432\* | 0.322\* | 0.551\*\* | 0.547\*\* | 0.461\* | 0.675\*\* |
| % Total Smoker | 0.452\* |  |  | 0.207 |  |  | 0.142 |  |  |
| % Male Smoker |  | 0.193 |  |  | -0.003 |  |  | 0.007 |  |
| % Female Smoker |  |  | 0.801\*\* |  |  | 0.738\*\* |  |  | 0.634\*\* |
| -------------------------------------------------------------------------------------------------------------------------------------- | | | | | | | | | |
| 1996 |  |  |  | 2004-05 Lung Cancer Death Rate | | | 2013 Lung Cancer Death Rate | | |
| Area Factors |  |  |  | 29 regions | | | 29 regions | | |
|  |  |  |  | Total | Men | Women | Total | Men | Women |
| GDP p. c. |  |  |  | 0.644\*\* | 0.688\*\* | 0.526\*\* | 0.232 | 0.186 | 0.295 |
| Cons. p.c. |  |  |  | 0.602\*\* | 0.654\*\* | 0.473\*\* | 0.225 | 0.18 | 0.285 |
| % Agri. Labor |  |  |  | -0.682\*\* | -0.646\*\* | -0.687\*\* | -0.267 | -0.171 | -0.443\*\* |
| % Ind. Labor |  |  |  | 0.826\*\* | 0.84\*\* | 0.733\*\* | 0.419\* | 0.349\* | 0.522\*\* |
| % Cons. Labor |  |  |  | 0.513\*\* | 0.487\*\* | 0.504\*\* | 0.378\* | 0.372\* | 0.349\* |
| % Serv. Labor |  |  |  | 0.619\*\* | 0.615\*\* | 0.577\*\* | 0.224 | 0.177 | 0.29 |
| Literacy |  |  |  | 0.624\*\* | 0.616\*\* | 0.57\*\* | 0.631\*\* | 0.606\*\* | 0.606\*\* |
| HS edu |  |  |  | 0.54\*\* | 0.509\*\* | 0.553\*\* | 0.168 | 0.085 | 0.327\* |
| Hsptl Bed |  |  |  | 0.421\* | 0.365\* | 0.488\*\* | 0.126 | 0.026 | 0.335\* |
| Med. Persn |  |  |  | 0.449\*\* | 0.389\* | 0.523\*\* | 0.051 | -0.056 | 0.278 |
| Waste Water p.c. |  |  |  | 0.583\*\* | 0.671\*\* | 0.397\* | 0.257 | 0.236 | 0.269 |
| Gas p.c. |  |  |  | 0.612\*\* | 0.616\*\* | 0.562\*\* | 0.24 | 0.164 | 0.382\* |
| Solid Waste p.c. |  |  |  | -0.34 | -0.363 | -0.266 | -0.233 | -0.248 | -0.176 |
| % Total Smoker |  |  |  | -0.12 |  |  | -0.019 |  |  |
| % Male Smoker |  |  |  |  | -0.077 |  |  | -0.092 |  |
| % Female Smoker |  |  |  |  |  | 0.393\* |  |  | 0.328\* |

Note: \* *p-value<0.05, \*\* p-value<0.01*

Supplement Figure 1. Province level administrative units and regions in China

Cited References

1. China State Statistical Bureau, editor. China Statistical Yearbook, 1997. Beijing: China Statistical Publishing House; 1997.

2. China State Statistical Bureau, editor. China Statistical Yearbook, 1986. Beijing: China Statistical Publishing House; 1986.

3. All China Marketing Research Co. Ltd. China historical county population census data with GIS maps (1953, 1964, 1982, 1990, 2000). China historical county population census data with GIS maps (1953, 1964, 1982, 1990, 2000): China Data Center; 2005.

4. CIESIN SEDAC. The China Administrative Regions GIS Data: 1:1M, County Level, 1990. 1996.

5. Agresti A. An Introduction to Categorical Data Analysis, Second Edition. Hoboken, NJ: Wiley; 2007.

6. SAS Institute. The GLIMMIX Procedure. SAS/STAT® 92 User’s Guide, Second Edition. Cary, NC2009. p 2250-5.

7. Cameron AC, Windmeijer FAG. R-Squared measures for count data regression models with applications to health-care utilization. J Bus Econ Stat **1996**;14(2):209-20 doi Doi 10.2307/1392433.

8. Han J, Zhu L, Kulldorff M, Hostovich S, Stinchcomb DG, Tatalovich Z*, et al.* Using Gini coefficient to determining optimal cluster reporting sizes for spatial scan statistics. International journal of health geographics **2016**;15(1):27 doi 10.1186/s12942-016-0056-6.