Growth, Pollution, and Life Expectancy: China from 1991-2012

Avraham Ebenstein, Maoyong Fan, Michael Greenstone, Guojun He, Peng Yin, Maigeng Zhou

Online Appendix

I. Mortality and Life Expectancy Data

Our measure of mortality is taken from the Disease Surveillance Point System (DSPS) of China. For a selected surveillance point (either a county or a city-district), the DSPS collects data on all deaths in hospital or at home. To represent national population and mortality trends, the DSPS adopts a multi-stage cluster population probability sampling method. The main objectives of the DSPS are to: (1) identify the number of deaths related to each disease category and provide basic mortality information about the deceased for public health officials; and (2) provide feedback to evaluate the impacts of the public health interventions. The DSPS initially covered 71 counties and city-districts in 29 provinces; this was expanded to 145 counties and city-districts in 31 provinces in 1990. The DSPS was overhauled following the SARS outbreak in 2003 and 161 counties and city-districts were designated in the system from 2004 to the present. Currently, the DSPS covers more than 81.5 million people or roughly 6 percent of the Chinese population. Our data period ranges from 1991 through 2012.[1] We have 102 DSPS monitor points that existed in the whole period and 84 with valid pollution measurements. It is worth noting, however, that the DSPS post-2004 defined as its coverage population the entire county's population, whereas that prior to 2000, each point covered roughly a third of the county's population.

Death rates at each DSPS location are calculated for 19 age groups: 0, 1, 2-5, 6-10, 11-15…, 75-80, 80-85 and older than 85. These age-specific rates are then used to create age-adjusted rates using the national age distribution for China in 2000. The age-adjusted rates are used to generate abridged life tables, which are in turn used to calculate life expectancy at each age using standard demographic techniques.

[1] Due to the large scale transition of data collection system in early 2000s, the mortality data in 2001, 2002, and 2003 are poorly collected and the China's CDC deems the data in those three years unusable.

II. Pollution Data

Assigning pollution accurately to DSPS locations has several challenging aspects. First, China's *Environmental Yearbooks* have incomplete readings for some of the stations, requiring us to interpolate missing data. Second, in limited instances, the data reported by the World Bank are not consistent with the reported reading in the *Environmental Yearbooks*. Third, we are forced in certain circumstances to choose between assigning pollution from a station which is closer to the DSPS location versus a station with more complete data. While in this project we follow the method outlined by Chen et al. (2013), in future work we plan to explore the sensitivity of our results to alternative decisions. In the remainder of this appendix, we briefly describe how we generated our pollution data, and assigned these data to DSPS locations.

Our pollution data are formed by combining several sources of Chinese air pollution, and is to our knowledge the most comprehensive data ever assembled for the period 1981-2012. These data were compiled through hand entry from Chinese language publications; most readings were taken from China's *Environmental Yearbook*s 1990-2012 and China's *Environmental Quality Annual Reports* 2001-2012, and verified against each other. In more recent years, the Chinese EPA provided us electronic copies of the data sources, which we used to verify our readings taken from the hard-copy publications.

Our data contain measures of several pollutants. The main pollutant we focus on is particulate matter, since this is considered the most harmful form of air pollution. The Chinese monitoring system reported Total Suspended Particulates (TSP) between 1981 and 2004 for a collection of cities, but discontinued reporting TSP in 2005. For 2003-2012, we have data on PM_{10}, particulate matter which is considered a more dangerous form of pollution. In our data, SO_2 concentrations have been reported consistently during our sample period (1990-2012). Concentrations of NO_X were reported for selected cities before 2001. Since 2001 all reporting is in terms of NO_2. In order to facilitate analysis over our entire period, such as in the first difference specification, we impute PM_{10} from recorded TSP by assuming 37% of TSP is PM_{10} (which is based on the average for periods in which both were recorded). Similarly, NO_2 is assumed to be assumed to be 75% NO_x.

The assignment of pollution to DSPS locations is equivalent to the manner described in Chen et al. (2013). In brief, we create a data set of pollution readings by monitoring station and year from China's *Environmental Yearbooks*. We then calculate the distance for each pair of

monitoring stations and DSPS locations. The distance between each of these stations and our 204 DSPS locations yields a full matrix of calculated distances. Our measure of air pollution for a DSPS location in a year is calculated as the weighted average of air pollution at each station, with the weights determined by the inverse of the distance between the two points, provided the station is less than 150 kilometers from the DSPS location. When a station had no valid TSP or PM_{10} reading for a particular year, it was assigned a zero weight for that year and did not enter into the calculation. For DSPS locations within 25 kilometers of a monitoring station, we simply use the reading from the closest station instead of the distance-weighted measure. For DSPS locations not within 150 kilometers of a monitoring station, they are not assigned pollution and are dropped from the sample.

Finally, at each DSPS location, the air pollution exposure was measured as the average TSP reading in all previous years when data are available. The assignment rule is to use *all* available data for each location retrospectively, and use TSP/PM_{10} conversions when necessary. For example, the reading for a DSPS location in 1991 is the average of TSP readings from 1981-1990 at the location. For a DSPS location in 2010, the reading would be calculated using all PM_{10} readings between 2003 and 2009, and then using TSP*.37 for 2002 and earlier as our imputed measure of PM_{10}. This creates a panel data set of year by DSPS location mortality readings with validated air pollution data averaged across all previous years.

III. Income and Demographic Data

We collected GDP and population data from China's statistical yearbooks for each DSPS location. For urban DSPS locations, the city-level GDP and population measures were taken from China's City Statistical Yearbooks. For rural DSPS locations, county-level GDP were taken from China's County Statistical Yearbooks, and Provincial Statistical Yearbooks when necessary. Per capita GDP is calculated as GDP divided by end-year population. We then adjust for inflation using China's national CPI (2010=100).

Our measure of GDP for 1990 and 2012 is generated using China's statistical records, which are comprehensive at the province-level but incomplete at the county or city level. In 2012, we are able to observe actual per capita GDP at the county level or city level for all but a few counties in Tibet. For these counties, we use province-level GDP per capita. In 1990, county-level

GDP is not available. As such, we impute per capita GDP in 1990 using the 2000 GDP per capita, and deflating by the province growth in GDP between 1990 and 2000.