Ruwan Thilakaratne

July 19, 2021

README for PHIRE data analysis scripts

All files located here: S:\Wildfire-Hoshiko\Wildfire - RThilakaratne\PHIRE\_Health\_Analysis\scripts

**Preliminaries**

* **RH greater than 100**: examines relative humidity (RH) values from NOAA dataset that are greater than 100%; examines spatial distribution and spatial correlation of these values with others, concluding that it is not unreasonable to impute these values to be 100%. Values >100% occur because RH is computed through spatiotemporal models, not directly measured
* **Reading old NOAA and Harvard data (2008-2009) :** Ana first provided NOAA and Harvard data as two datasets, for 2010-2016 and for California specifically. We later needed to add 2008 and 2009 data – she had these as daily files for the entire US. I created this script to process those files into single California-specific files, which are later merged with the analytic dataset in S:\Wildfire-Hoshiko\PHIRE OSHPD DATA\Analytic datasets\Merging in covariates.rds
* **PM25 models:** characterizing basic spatial and temporal distributions of Harvard PM2.5 data; creating ZIP3-based residuals (note both ZIP3 and climate zone-based residuals are generated in the more recent dataset assembly script S:\Wildfire-Hoshiko\PHIRE OSHPD DATA\Analytic datasets\Merging in covariates.rds, with the latter residuals chosen for model building)
* **Improving run time for random intercept models:** by default, random intercept models run using the glmer function from package lme4 use the LaPLace approximation to fit the model intercepts. PIRLS (partially iterated reweighted least squares) is another much faster option, though it doesn’t achieve exactly the correct parameters values. However, this script shows the difference is so small as to be negligible for our purposes, so the latter approach is used
* **Examining time series at ICD cross for all main outcomes:** graphs cumulative (across zips, i.e. CA total) daily time series of the 4 main outcomes for 2014, 2015, and 2016 to visually examine how ICD coding switch on October 1 2015 affects time series contiguity
* **ED and hosp data time series:** state-wide (summed over zips) time series over study period for all major outcomes, juxtaposing ED vs HA. Interestingly, ED rises a lot over time but hosp doesn’t. HA has clear weekday vs. weekend patterning.
* **Checking dispersion parameters in models (ED and PDD):** runs final models and estimates dispersion parameter as residual deviance/degrees of freedom. 1 = Poisson dispersion, >1 = over-dispersed, <1 = underdispersed.Found that all dispersion parameters were near 1 (main concern is overdispersion which would underestimate standard errors)

**Model building**

* Script titles should be relatively self-explanatory. The more involved model building was done for asthma ED visits, as that was the first outcome we pursued. Subsequently, the same model structure was replicated for other outcomes and for hospitalizations, except for differences in the lag period – thus, scripts for the other outcomes and hospitalizations center on modeling different distributed lags and selecting the one with best fit.
* First model built was for asthma ED, before expanding to other outcomes. “Model building – candidate models” contains all asthma ED models that were run in the process of selecting a best model. Several aspects of the model were prespecified and the same across candidate models: heat index as a spline with 3 DOF; day-of-week; Poisson distribution assumed. Components that varied:
  + In some models, PM is left in its original form, and in others it’s detrended. We ended up using residuals detrended by climate zone instead of ZIP3, because climate zone residuals BIC was better than ZIP3 residuals BIC.
  + We varied time spline df/year in models using PM residuals to see if PM effect was sensitive to this parameter. It had minimal impacts. However, in order to explain outcome variability for predictions for burden calcs, we needed some kind of time spline. So, we tired selecting the best DF/year based on BIC for asthma (to start), but found that BIC continued to decrease even with an unreasonable number of DF. Therefore, we chose to select the DF *a priori* at 4df/year, as this models trends of 3 months or more, which is typical in time series studies (will cite a paper – see manuscript draft). So, DF by best BIC is not pursued for other outcomes (CV, resp, COPD, and hospitalization versions of the outcomes), and instead prespecified at 4df/year.
  + We needed to account for spatial confounding because this analysis spans such a large geographic range. We used Moran’s I, a measure of spatial autocorrelation, to proxy spatial confounding (smaller Moran’s I is better). Moran’s I is computed for several spatial modeling choices – ZIP3 fixed effect, Healthy Places Index (HPI), etc.
    - Upon discovering the late lag effect (a product of spatiotemporal confounding due to the modeled nature of the PM2.5 data) and that a zip code-level term was the only way to eliminate it (namely, zip code random intercept – a fixed effect does the same thing but doesn’t run because R can’t handle that many separate terms in the model), Moran’s I was calculated for this model and compared to other spatial modeling options, and was found to be superior.
* **Alternative estimation for zip code fixed effect:** estimates RR for asthma ED visits using same model but alternative method (with different assumptions) – random effects meta analysis. Estimates log-relative risk for all ZIP3’s and then combines them using random effects meta analysis to generate a summary effect and confidence interval. This assumes the true effect is different in different ZIP3’s, and the summary effect is an average of these effects, though isn’t actually an observed effect itself (just an average of other effects). The model we went with is essentially a fixed effect model with respect to PM2.5, that estimates a single PM2.5 RR
* **Troubleshooting late lag effect**: investigates late lag effect by identifying affected ZIP3’s; maps the presence of effect by ZIP3; runs negative control experiment in which PM2.5 residuals are put in random order (should produce null effects but for ZIP3 + offset models, lag0 and last lag risks remain

**Tables and figures**

* **Stratified models:** builds all PM health effects models (ED and HA, overall and all subgroups), and stores them in S:\Wildfire-Hoshiko\Wildfire - RThilakaratne\PHIRE\_Health\_Analysis\output\Risk estimate models. Also stores results by type of visit (ED or HA) and outcome (e.g. ed\_allcause\_cv\_estimates.rds) and cumulative results for each type of visit (e.g. ed\_all\_results.rds has all ED effect estimates). Performs relabeling so they’re ready for presentation in plots and tables (see next scripts).
* **Figure – ED risk estimates:** generates figure out of overall and subgroup risk estimates (point estimates and 95%CI’s) for ED outcomes using ggplot (drawing from estimates stored here: RThilakaratne\PHIRE\_Health\_Analysis\output\Risk estimate models\ ed\_all\_results.rds ).
  + **This has been dated by the forest plots created by Amrita (Ana’s colleague).** The Native American interval throws off the y axis scale due to its width – forestplot can easily truncate the interval and keep the smaller scale, so we opted for it.
* **Figure – HA risk estimates:** ditto but for HAs
* **SI Table – ED risk estimates:** R markdown file that creates table from ED risk estimates, using kableExtra package. Table was then exported to HTML, pasted into Excel, reformatted then finally pasted into Word
* **SI Table – HA risk estimates:** ditto but for HAs
* **Sensitivity analyses:** runs sensitivity analyses varying degrees of freedom in time spline (5 and 6 per year instead of original 4 per year) for overall outcome models (overall ED and HA outcomes, not subgroups), and stores models and estimates in “S:\Wildfire-Hoshiko\Wildfire - RThilakaratne\PHIRE\_Health\_Analysis\output\Risk estimate models\Sens analysis - spline dof”. These will be turned into a figure by Amrita.