Chapter 1

Exposure-Response Analysis Using LSTM Model

This chapter focuses on methods and models for obtaining risk estimates from time series data and exploring the sensitivity of those estimates to modeling approaches. The proposed model first extracts features from multiple air pollutants time series with distributed lags through LSTM network.

1.1 Exposure-Response Coefficient in GAM

1.1.1 Traditional Exposure-Response Formulation

In general air pollution epidemiology studies, one of the important tasks for Generalized Additive Model (GAM) is to analyze the relationship between air pollution exposure and health consequences. Quantitative evaluation of the adverse impacts from air pollution on public health can facilitate future regulation of air pollutants emission, environmental policy making, development of social healthcare, etc. In a standard GAM used for short-term air pollution-related health risk assessment, its main component is the air pollutant predictor. Besides, natural cubic regression spline functions are used as smoothers for temperature and time to accommodate the measured covaritate and unmeasured confounders such as seasonal variations and other time-dependent risk factors that may influence the health outcome of interest. This health risk assessment model is rewritten in Eqn. (1.1) to highlight the different components and the exposure-response relation.

$$\log(\mu_t) = \beta_0 + \beta_1\chi_t + \beta_2 DOW_t + \underbrace{\mathrm{NS}(T, df)_t + \mathrm{NS}(Time, df)_t}_{\text{Smoothers}} \tag{1.1}$$

In Equn. (1.1), β_1 is the parameter of interest that reflects the air pollution exposure and health response relation. It means the rate at which the logarithm of health consequence μ_t (such as daily mortality) changes with the increase of concentration level of air pollutant χ_t. This coefficient is also the sensitivity that a specific public health issue responds to people's exposure to some air pollution.

1.1.2 Reformulation with Slepian Filter

The two smoothers of natural cubic regression splines function for temperature and time in Eqn. (1.1) are equivalent to the formulation using same orthonormal bases in the following Eqn. (1.2),

$$\log(\mu_t) = \beta_0 + \beta_1\chi_t + \beta_2 DOW_t + \underbrace{\sum_{i=1}^{n} \zeta_i Base_i}_{\text{Smoothers}} \tag{1.2}$$

where $Base_i$ and ζ_i are the basis of spline function and the corresponding coefficient. The above formulation provides a general framework for accommodating various confounding factors. However, in air pollution epidemiology

Introduce the high-pass filter a Slepian high-pass projection filter and a Slepian projection smoother [1, 2]

The bases derived from the smoother function effectively divide the response space into two subspaces: one spanned by the smoother bases, and another orthogonal to this smoother base-spanned subspace. Given this decomposition, solving the linear model in Equation (1.2) can be simplified by pre-filtering the response and air pollutant variables. This reduces the model to a simpler two-predictor form, as shown in Equation (1.3). In (1.3), $death^f(t)$, $PM10^f(t)$ and $O3^f(t)$ represent the time series data for daily mortality counts, PM10 daily mean, and ozone daily mean, respectively, after being filtered by the smoother bases.

$$\log(death^f(t)) = \beta_0 + \beta_1 PM10^f(t) + \beta_2 O3^f(t) \tag{1.3}$$

In environmental epidemiology studies focused on short-term risk estimation, the time smoother is designed to capture the long-term relationship between health risks (the response) and all other unmeasured confounders in a long-term sense. While the selection of smoother functions is beyond the scope of this paper, it is important to note that a conventional approach often involves using natural cubic splines as the time smoother. These splines are piecewise third-degree polynomial functions. However, this conventional choice may not effectively separate long-term effects from short-term fluctuations.

To address this limitation, this model employs a Slepian smoother. The Slepian smoother allows for a clearer distinction between short-term and long-term effects. When using the Slepian smoother, both the response data (death counts) and the air pollutant data are preprocessed with a high-pass filter. This filter blocks low-frequency patterns in the time series data, effectively excluding long-term effects and focusing the analysis on short-term risks.

The difference between the health outcomes using different air pollution predictors is used to quantitatively assess the risk of exposure to specific air pollutants as follows.

A group of selected air pollutants $\mathbf{\Gamma}$ is first used as predictors by the GAM to assess some health outcome as in Eqn. (1.4),

$$\log(\mu_t) = F_{\mathrm{GAM}}(T, \mathbf{\Gamma}) \tag{1.4}$$

where $F_{\mathrm{GAM}}(\cdot)$ represents the fitting function of the GAM.

Then the air pollutant of interest χ is taken away from $\mathbf{\Gamma}$ and another GAM is used to fit the health outcome as Eqn. (1.5),

$$\log(\mu_t^-) = F_{\mathrm{GAM}}^-(T, \mathbf{\Gamma}^-) \tag{1.5}$$

where μ_t^-, $F_{\mathrm{GAM}}^-(\cdot)$ and $\mathbf{\Gamma}^-$ are the health outcome, fitting function of the GAM and the group of air pollutants without χ. Note that the difference between Eqn. (1.4) and Eqn. (1.5) is the kinds of air pollutants used in the GAMs.

The difference between the health outcome between μ_t of Eqn. (1.4) and μ_t^- of

Eqn. (1.5) is shown as y in Eqn. (1.6).

$$y = \exp[F_{\mathrm{GAM}}(T, \mathbf{\Gamma})] - \exp[F^{-}_{\mathrm{GAM}}(T, \mathbf{\Gamma}^{-})] \tag{1.6}$$

A linear regression model as Eqn. (1.7) is finally used to evaluate the impacts of χ on the variation of the assessed health outcome,

$$y = \beta_{\mathrm{GAM}} \times \chi \tag{1.7}$$

where β_{GAM} is the expose-response coefficient that represents the impacts of air pollutant χ on the variation of assessed health outcome. This coefficient provides a quantitative risk assessment measure of the impacts from exposure to χ .

1.2 Exposure-Response with LSTM Model

Following the exposure-response formulation in GAM, preliminary exploration of the quantitative association between specific health outcome and air pollutant of interest is made using the proposed Long Short-Term Memory (LSTM)-based air pollution-related health assessment model. The input difference to the LSTM model is specifically designed to emphasize the air pollutant for investigation.

First, a group of selected air pollutants is used as the input to assess their impacts on some specific health outcome as in Eqn. (1.8),

$$\nu_t = F_{\mathrm{LSTM}}(T, \mathbf{\Gamma}) \tag{1.8}$$

where ν_t is the health outcome to be assessed, $F_{\mathrm{LSTM}}(\dot{})$ represents the fitting function of the LSTM model, and $\mathbf{\Gamma}$ represents the group of selected air pollutants.

Then the air pollutant for investigation is taken away and another LSTM model is trained to assess the health outcome as Eqn. (1.9),

$$\nu_t^- = F_{\mathrm{LSTM}}^-(T, \mathbf{\Gamma}^-) \tag{1.9}$$

where ν_t^-, GAM$^-$ and $\mathbf{\Gamma}^-$ are the health outcome, fitting function of the LSTM model and the group of air pollutants without the air pollutant of interest. Note that all parameters of the LSTM network and the estimation layers keep the same for Eqn. (1.8) and Eqn. (1.9) and the only difference is the types of air pollutants input to the model.

Finally, a linear regression model as Eqn. (1.10) is formulated to evaluate the impacts of air pollutant of interest on the variation of the assessed health outcome,

$$y = \beta_{\mathrm{LSTM}} \times \chi \tag{1.10}$$

where $y = \nu_t - \nu_t^-$ is the difference of health outcome between Eqn. (1.8) and Eqn. (1.9), and χ is the air pollutant of interest. β_{LSTM} is the expose-response coefficient that represents the impacts of air pollutant x on the variation of assessed health outcome.

1.3 Data Preparation

Besides Chicago, data of 10 populous cities are selected from the 108 U.S. cities in the National Morbidity and Mortality Air Pollution Study (NMMAPS) database, and are prepared for the following numerical experiments. These cities include Los Angeles, New York, Dallas/Fort Worth, Houston, San Diego, Miami, San Bernardino, San José, Riverside, and Philadelphia. The reasons for selecting these cities and preparation of the data are as follows.

1) The population of each selected city was more than 1.3 million by the year 2000, which can facilitate the statistical analysis of the air pollution-related health risk.

2) Considering that not all cities in the NMMAPS database have complete measurement of the air pollutants of interest, these cities with relatively less missing data are selected for analysis.

3) All-cause non-accidental daily mortality is evaluated with particulate matter less than or equal to 10 micrometers (PM_{10}) and ground ozone (O_3) as the air pollutants of interest. For each city, the daily non-accidental mortality data for three age categories (under 65, 65 to 74, and above 75) are aggregated as the daily morality observation, similar to the preparation of Chicago dataset.

4) For each city, the original air pollution data were detrended by subtracting a 365-day moving average from them. The air pollution data are restored to the measured values by adding the 365-day moving average in order to facilitate the following experiments.

5) Missing values of the air pollutants and the health outcome for each city are interpolated using a second-order polynomial function, and the outliers are dealt with as in experiments of previous chapters.

6) Most PM_{10} concentration levels in the NMMAPS database were recorded every six days while the O_3 data were collected on a daily basis. Taking into account the formulations of both the LSTM model and the GAM, all these data are reorganized on a 6-day interval, with which the original 14 years data are regrouped

into a time series of about 852 periods.

7) As there exist significant differences between weekdays and weekends for both the air pollution concentration level and the corresponding health consequence resulting from people's activities, all categories of data for 10 cities are processed as Eqn. (1.11) to remove the differences.

$$\gamma^{'}(t) = \gamma(t) - \overline{\gamma} \tag{1.11}$$

In Eqn (1.11), $\gamma(t)$ is the original data at period t before processing, such as temperature, an air pollutant and some health outcome. $\gamma^{'}(t)$ is the corresponding data after being processed. $\overline{\gamma}$ represents the mean value of $\gamma(t)$ of the same day of week across all periods. For example, if t is Monday, then $\overline{\gamma}$ is the mean value of $\gamma(t)$s on all Mondays in the dataset.

8) When applying the filters to the response variable, daily death counts, a challenge arises because fitting negative values to a Poisson distribution is not feasible. To address this issue, we add a constant value M to all filtered daily death counts. We then offset this value in the GAM model by deducing the constant value M back to the predictive death counts.

1.4 Results and Discussions

1.4.1 Experimental Results

Experiments on Chicago Dataset

Figure 1.1 illustrates the LSTM-based slopes calculated using the Chicago dataset. These slopes are computed across different looking-back step settings, ranging from 1 to 8, using the predictive daily death counts from both the 'full' LSTM model and the 'lack' LSTM model. The largest slope (0.035) is observed in the LSTM model with a looking-back step of 1, while the smallest slope (-0.006) occurs at a looking-back step of 5. For models with looking-back steps greater than 4, the absolute values of their LSTM-based slopes are all smaller than 0.01. In contrast, models with looking-back steps of 4 or fewer generally exhibit LSTM-based slopes larger than 0.01, except for the model with a looking-back step of 3.

Comparison of Slopes for GAM and LSTM models

With the preprocessed datasets for each city, we calculated the LSTM-based mortality association coefficients. Figure 1.2 shows the slopes for each city across different look-back steps in the LSTM models. For each city, no clear trend (increasing or decreasing) in LSTM-based slopes with the increase in look-back steps is observed. Overall, New York City exhibits larger LSTM-based slopes compared to other cities, whereas Los Angeles tends to have smaller slopes.

Next, we calculated the LSTM-based mortality association coefficients without

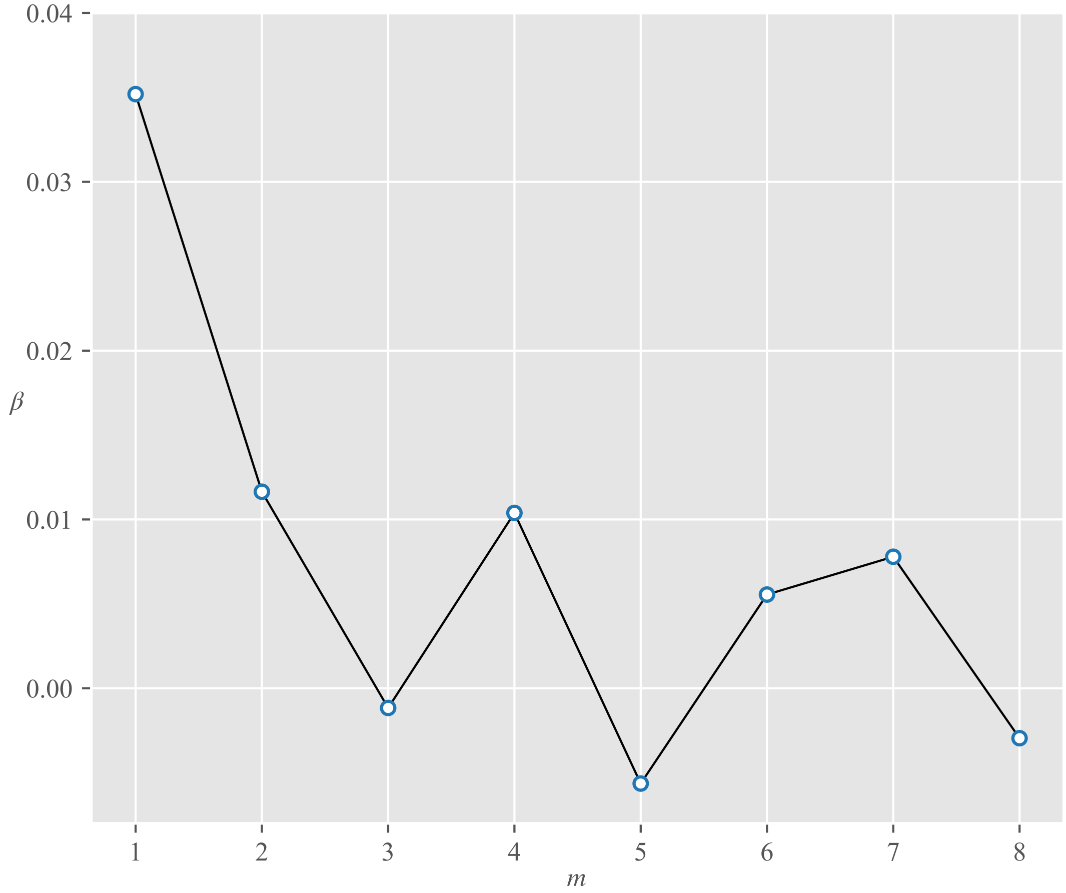

Figure 1.1: LSTM based PM10-Mortality association coefficients(LSTM based Slopes) with different looking back steps with Chicago dataset.

setting a seed, meaning that the LSTM model would predict daily death counts differently each time we train the model. Figure 1.3 illustrates the distributions of LSTM-based slopes calculated from 30 training iterations. Overall, New York City shows the largest variance in LSTM-based slopes across all look-back steps, compared to the other cities. Aside from New York, the magnitude of the LSTM-based association coefficients is generally below 0.025. Again, no consistent trend in the slopes with increasing look-back steps is apparent across the cities, for both the mean values and the variance.

Based on the LSTM-based slope statistics for the 10 cities, it appears that the

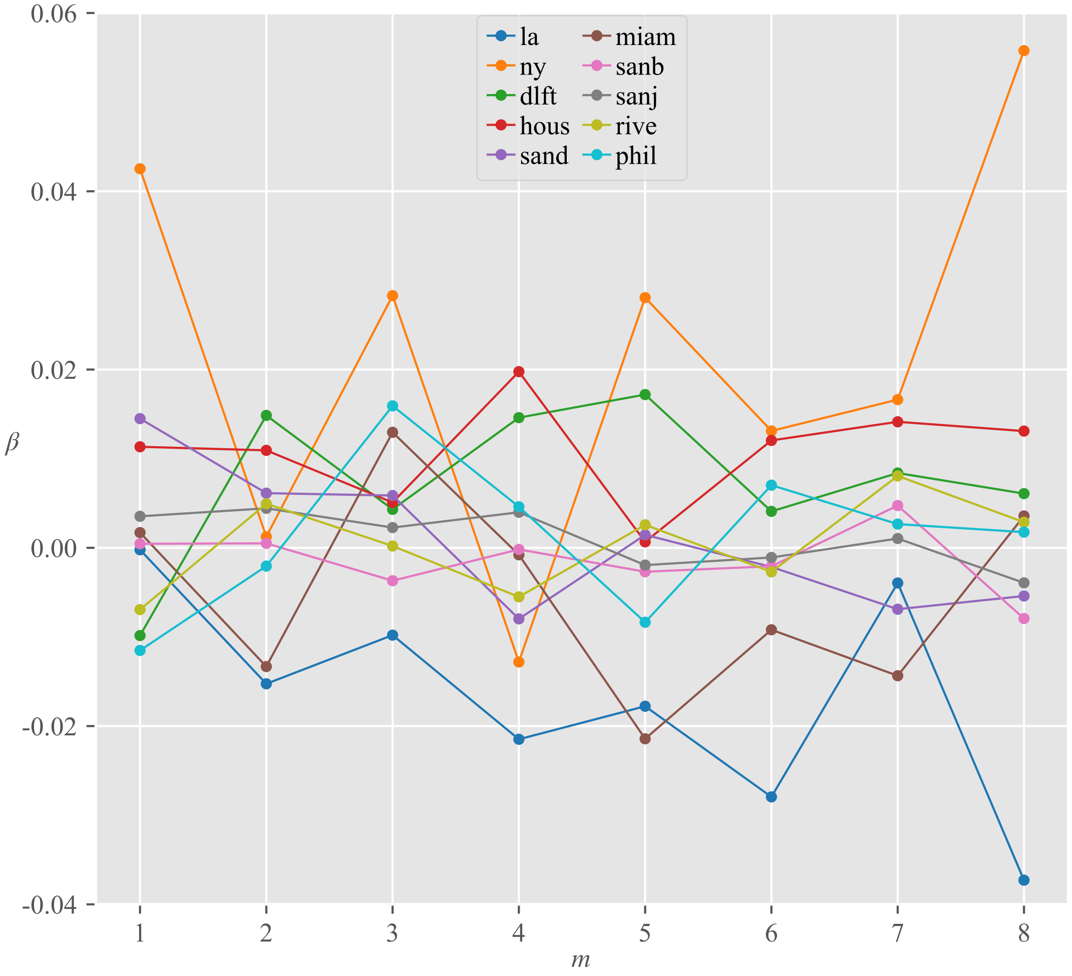

Figure 1.2: LSTM based PM10-Mortality association coefficients(LSTM based Slopes) with different looking back steps for 10 cities in the US.

look-back steps in the LSTM models do not significantly influence the final association coefficients. Thus, we set the look-back steps to 5 for subsequent comparisons. Since New York City and Los Angeles exhibit outlying behavior in terms of LSTM-based slopes, we exclude them and focus on the remaining 8 cities. Figure 1.5 compares both LSTM-based and GAM-based slopes for these 8 cities. A clear linearity is observed between the two types of association coefficients. Dallas/Fort Worth has the largest LSTM-based association coefficient (0.017) and the largest GAM-based association coefficient (0.052), while Miami shows the smallest LSTM-based association coefficient (-0.021) and the smallest GAM-based association coefficient (-0.007). This result demonstrates that LSTM models provide similar results to GAM models in terms of association coefficients for health risk assessment.

The analysis of LSTM-based PM10-mortality association coefficients across multiple cities and look-back steps yielded several insights. Firstly, the lack of a clear trend with increasing look-back steps suggests that the chosen temporal window in the LSTM model does not significantly affect the estimated slopes.

New York City stood out for having both larger slopes and greater variability compared to other cities. This may indicate a more dynamic relationship between PM10 exposure and mortality in New York, potentially driven by the city's unique environmental, demographic, and urban characteristics. In contrast, Los Angeles consistently exhibited smaller slopes, suggesting a weaker association between air pollution and mortality, which might be attributed to differing pollution sources, weather patterns, or public health responses in the region.

Interestingly, the comparison between LSTM-based and GAM-based slopes revealed a strong linearity, underscoring the robustness of these models in estimating health risks related to air pollution. Although LSTM models are more complex and can capture non-linear relationships, the similarity in results to the simpler GAM model suggests that, for this dataset, the LSTM model show ability to give quantitative association relationship between a certain air pollutant and the mortality.

Future work could explore other factors, such as meteorological variables or demographic differences, to refine these models further. Additionally, further investigation into why New York and Los Angeles differ so markedly from other cities could provide critical insights into region-specific health policies and interventions.

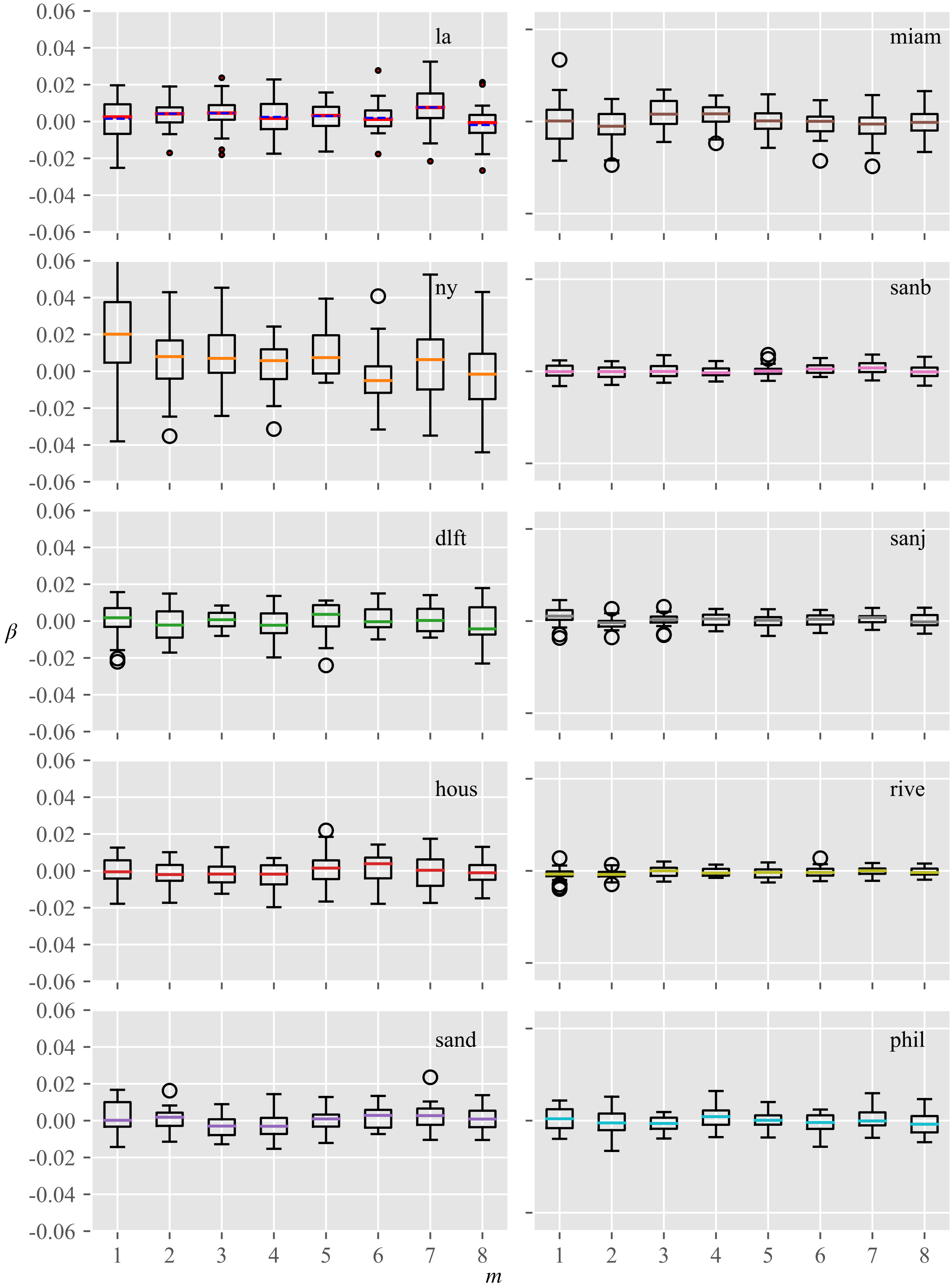

Figure 1.3: LSTM based PM10-Mortality association coefficients(LSTM based Slopes) with different looking back steps for 10 cities in the US. The distributions are represented in the boxplots for 30 times of training with LSTM models. (la: Los Angeles, ny: New York, dlft: Dallas/Fort Worth, hous: Houston, sand: San Diego, miam: Miami, sanb:r San Bernardino, sanj: San José, rive: Riverside, phil: Philadelphia)

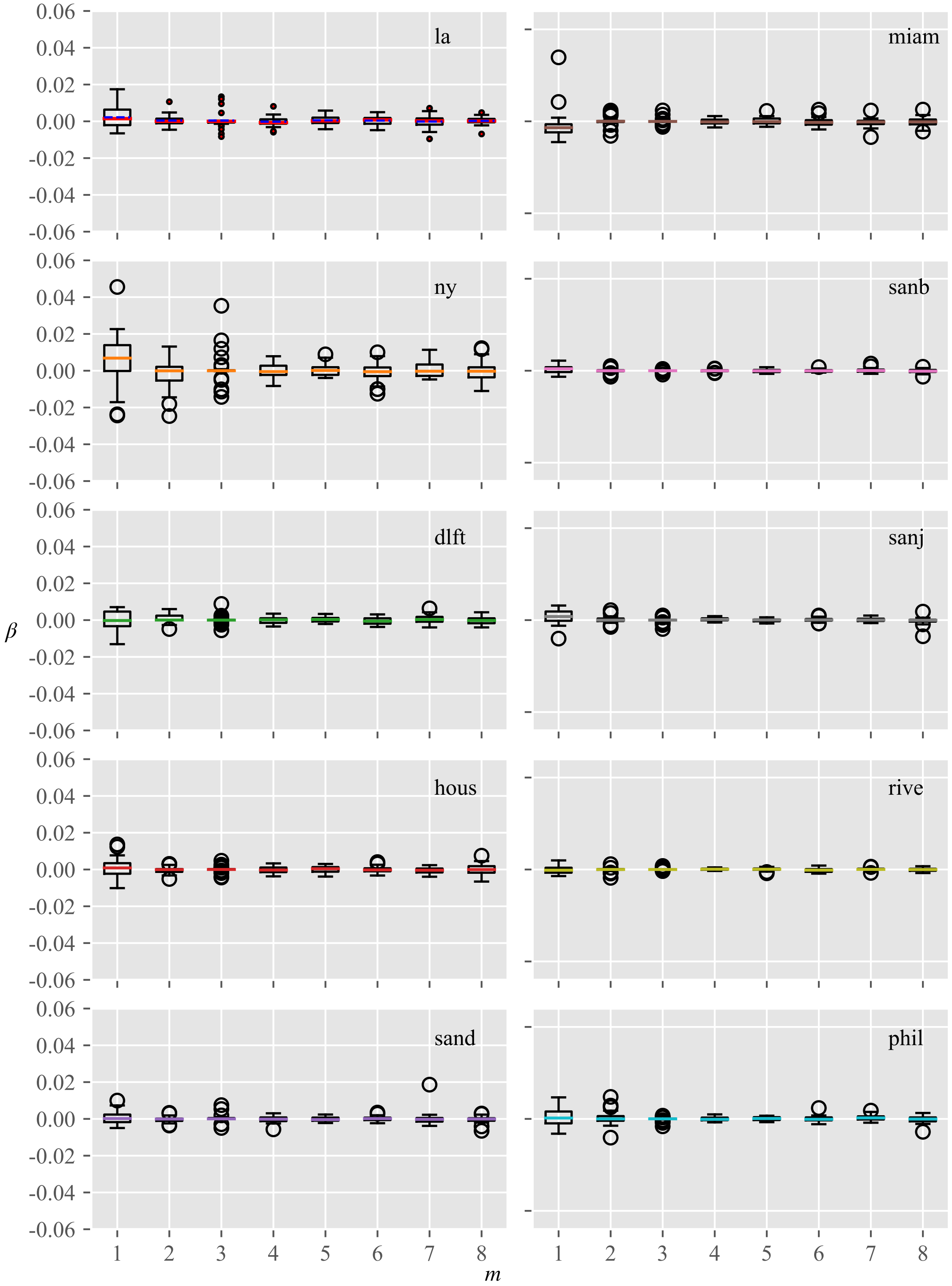

Figure 1.4: LSTM based PM10-Mortality association coefficients(LSTM based Slopes) with different looking back steps for 10 cities in the US. The distributions are represented in the boxplots for 30 times of training with LSTM models. (la: Los Angeles, ny: New York, dlft: Dallas/Fort Worth, hous: Houston, sand: San Diego, miam: Miami, sanb:r San Bernardino, sanj: San José, rive: Riverside, phil: Philadelphia)

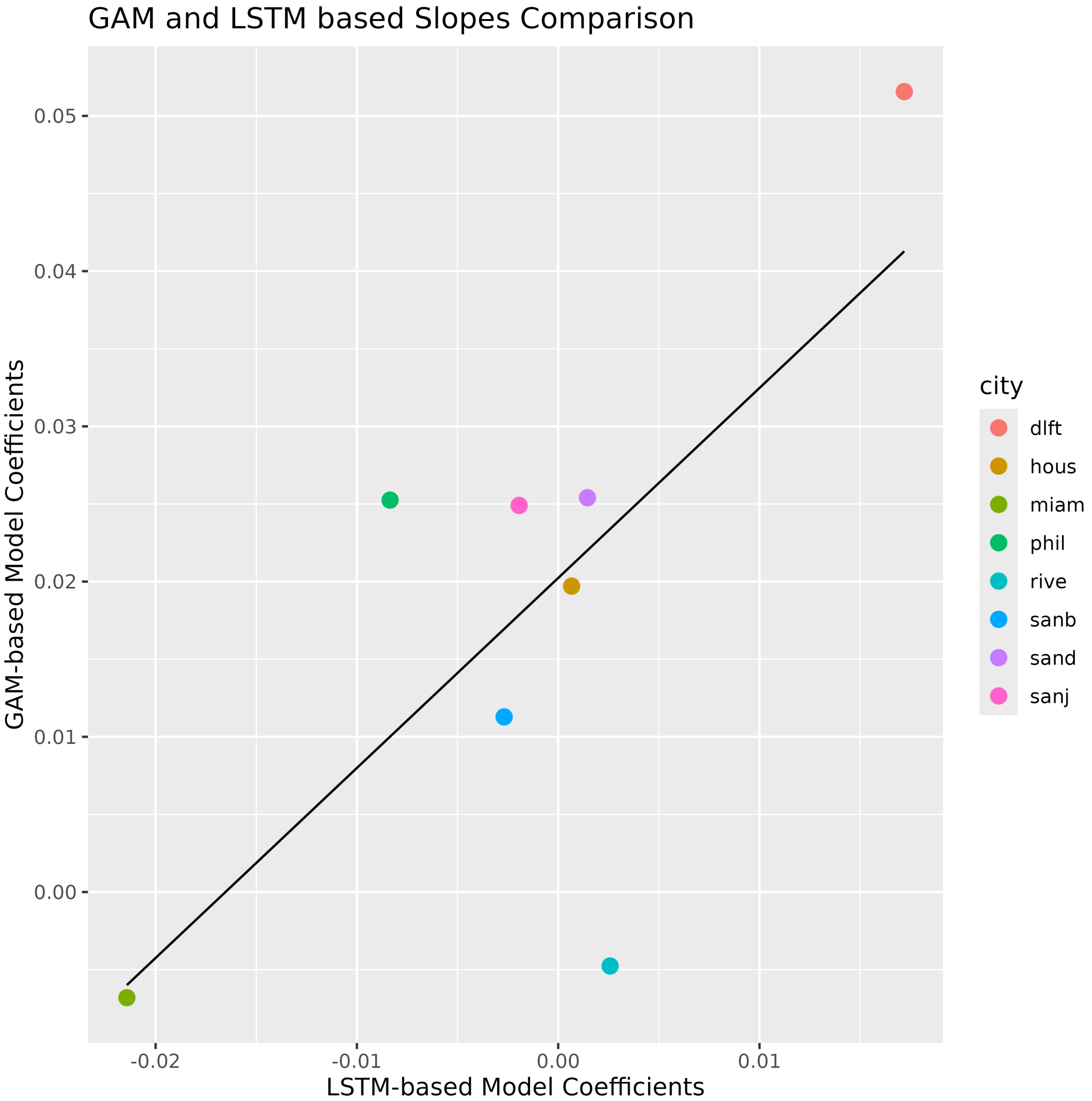

Figure 1.5: LSTM VS. GAM based association coefficients in 8 populous cities. The look back step for LSTM models is 5.

Bibliography

[1] Wesley S. Burr. *Air pollution and health: Time series tools and analysis.* PhD thesis, Queen's University (Canada), 2012.

[2] Wesley S. Burr. tsinterp: a time series interpolation package for R. https://github.com/wesleyburr/tsinterp, 2022.

APPENDICES

Functions Used in LSTM network

The sigmoid, softmax and tanh functions used in the LSTM model is as Eqns. (12-14).

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{12}$$

$$SF(i) = \frac{e^{i}}{\sum\limits_{j} e^{j}} \tag{13}$$

$$\tanh(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}} \tag{14}$$