Environmental Exposures, Population across Locations and Time

Fan Wang

2022-12-31

Contents

1 Time, Location, Population, and Environmental Exposures

Go to the **RMD**, **R**, **PDF**, or **HTML** version of this file. Go back to fan's REconTools research support package, R4Econ examples page, PkgTestR packaging guide, or Stat4Econ course page.

1.1 Population and Locations

We consider now the relationship between population and location during some span of time. We are not modeling why individuals move across locations, but are modeling the distributions of individuals across locations during some span of time.

The discussions below and statistics to be computed do not require that we observe movements across locations. The idea here is to provide a statistical framework for computing environmental exposure distributions that can accommodate a range of data inputs. These data inputs might be tracing the movements of surveyed household panels across locations, or the inputs might be from low-frequency cross-sectional aggregated census data.

1.1.1 Time, Socio-demographic Groups, Path, and Locations

We define several variables related to time, socio-demographics groups, locational-paths, and locations.

First, let us define y, T_y, and N:

- y: Some frame of time under consideration, 1 year for example.
- T_y: Number of sub-periods of time, where y is for example a particular year, and T_y might be the integer number of months or days within that year.
  - Sub-period time index: $t \in \mathcal{T}_y \equiv \{1, \ldots, T_y\}$.

- N: Number of intersectional socio-demographic groups based on observable characteristics that are invariant at least within period y.
  - Individual group index: $n \in \mathcal{N} \equiv \{1, \ldots, N\}$.

Second, let us define O_y and M:

- O_y: Number of potential paths across the M locations during period y for the sample/population we are studying.
  - Path index: $o \in \mathcal{O}_y \equiv \{1, \ldots, O_y\}$.
- M: Number of locations where ambient exposure variations come from.
  - Location index: $m \in \mathcal{M} \equiv \{1, \ldots, M\}$.

1.1.2 Share of Individuals across Socio-demographic Groups and Locational Paths

In this section, we define two matrixes, a matrix that defines the observed locational-paths, $\mathbf{Q}_y$, and a matrix that defines the distribution of socio-demographic groups across locational-paths, $\mathbf{P}_y$, both of these are period y specific.

The two matrixes have the same number of rows, which is he number of observed locational-paths O, which, as we will explain more later, equals the number of locations M if individuals do not move within period y.

First, we define let us define locational-path vectors and matrix:

- $\mathbf{q}_{o,y} = \left(m_{y,t=1}(o), m_{y,t=2}(o), \ldots, m_{y,t=T_y-1}(o), m_{y,t=T_y}(o)\right)$.
  - $\mathbf{q}_{o,y}$ is a 1 by T_y vector that tracks the location covered by path o at each point in time within $\mathcal{T}_\dagger$.
  - If individuals within period y do not move at all, then that means: $m_{y,t=1}(o) = m_{y,t'}(o)$, for all $t' \in \mathcal{T}_\dagger$. This could look like for example: $\mathbf{q}_{o=2,y} = (2, 2, \ldots, 2, 2)$, where path $o = 2$ is simply for location $m = 2$.
  - Given what is stated above, if individuals within period y do not move at all, $M = O$.
- $\mathbf{Q}_y = (\mathbf{q}_{o=1,y}, \mathbf{q}_{o=1,y}, \ldots, \mathbf{q}_{o=O-1,y}, \mathbf{q}_{o=O,y})$:
  - $\mathbf{Q}_y$ is a $(O \times T_y)$ matrix with O rows and T_y columns that tracks the locations covered by all $\mathcal{O}$ paths at each point in time within $\mathcal{T}_\dagger$.
  - If individuals in period y do not move at all, then $M = O$, and the number of rows of matrix $\mathbf{Q}_y$ would be equivalent to the number of locations. Furthermore, the $m_{t,y}$ values across columns in each row would be identical.

Second, consider the mass experiencing ambient exposures as summing to 1:

- $p_y(n)$: This is fraction of individuals in socio-demographic group n in overall time period y.
  - $\sum_{n=1}^{N} p_y(n) = 1$.
- $p_y(n, o)$: This is the fraction of individuals in socio-demographic group n and with locational-path o during time period y.
  - $\sum_{o=1}^{O} p_y(o, n) = p_y(n)$: All members of the group n experience one of the O locational-paths.
  - $p_y(o, n) = 0$ or $p_y(o, n) > 0$ for all $o \in \mathcal{O}$: There might be some paths that members of group n do not experience.
- $\mathbf{P}_y$ matrix:
  - $\mathbf{P}_y$ is a $(O \times N)$ matrix with O rows and N columns.
  - $p_y(n, o)$ is the fraction in the n^{th} row and o^{th} column.
  - $\sum_{o=1}^{O} \sum_{n=1}^{N} p_y(o, n) = 1$: total mass in the matrix sums to 1. The matrix accounts for the distribution of individuals of varying socio-demographic groups across locational-paths during period y.
  - Again, if individuals in period y do not move at all, then $M = O$, and the matrix $\mathbf{P}_y$ would accounts for the distribution of individuals of varying socio-demographic groups across locations during period y.

1.2 Ambient Exposure across Location and Space

So far, we have discussed the movements of individuals across locations and time, tracked by matrix $\mathbf{Q}_y$, and also the distribution of mass across socio-demographic groups and locational-paths, tracked by matrix $\mathbf{P}_y$. Now, we consider environmental exposures. Environmental exposures, including temperature, rainfall, pollution, and other components, can vary across time and space.

1.2.1 Location- and Time-specific Exposure

Let time be continuous and indexed by τ, and as before, there are M locations. Define $z_m(\tau)$ as the ambient exposure at location m at some moment of time τ. For example, within some bound of time t to $t+1$, and at location m, total exposure experienced by someone at location m is:

$$\int_t^{t+1} z_m(\tau) d\tau.$$

We convert the time-series of exposure to a distribution, we simply resort the exposures from the lowest to the highest and construct the corresponding cumulative function, with equal weights for each moment in time. Specifically, let us define the distribution of exposure between time t and $t+1$ at location M as:

$$F_{m,t}\left(Z \leq \widehat{z}\right) = \int_t^{t+1} \mathbf{1}\left\{z_m(\tau) \leq \widehat{z}\right\} d\tau,$$

where $\mathbf{1}$ is the indicator function.

1.2.2 Path-specific Exposure

The individuals we study might or might not move across the M locations during period y. To accommodate for the possibility that they do, we consider locational-path o-specific exposure distributions. This is needed because location m-specific or m'-specific exposure distributions during period y would not be the distribution that individuals who make a single move from location m to m' at some t during period y faces. The individuals who move will experience exposure at both location m and m' for different sub-segments of time within period y.

Specifically, the distribution of exposure along each locational-path o from $t=1$ to $t=T_y$ can be thought of as a mixture distribution, with equal weights $(\frac{1}{T_y})$ for each t time segment, which is the frequency over which potential movement data is observed. Specific, the locational-path o-specific cumulative exposure function during period y is:

$$F_{o,y}\left(Z \leq \widehat{z}\right) = \frac{1}{T_y} \sum_{t=1}^{T_y} \left(\sum_{\widehat{m}=1}^{M} F_{\widehat{m},t}\left(Z \leq \widehat{z}\right) \mathbf{1}_{\left\{m_{y,t}(o)=\widehat{m}\right\}} \right).$$

In our analysis here, individuals' exposures are defined by their locational-path and all moments of within-individual exposure distributions within period y are based on $F_{o,y}$. For example, based on the locational-path o-specific cumulative exposure function, we can compute any percentiles of interest via inversion. Specifically, the τ^{th} percentile of locational path o-specific distribution is:

$$Q_{o,y}\left(\tau\right) = F_{o,y}^{-1}\left(\tau\right).$$

We can also compute mean, standard deviation, and other statistics based on the $F_{o,y}(Z)$ cumulative distribution.

1.2.3 Exposure Data Discretization

In the last section, we considered τ as continuous, but any observed data will come at discretized frequencies. Higher frequency data will provide more precisely measured within locational-path individual-specific exposure moments, but sometimes only lower frequency data is available. We consider below differing empirical scenarios.

First, given our prior discussion, there are three layers of time:

1. y: the overall period within which we are consideration exposure distributions.
2. $t \in \mathcal{T}_y \equiv \{1, \ldots, T_y\}$: the unit of time over which data on movement/mobility of individuals across space is available.
3. τ vs d:
    - τ: The continuous unit of time over which environmental information in theory exists. Empirically, data is not observed continuously, but at some finite-time interval.
    - d: Empirically, $z_m(\tau)$ is measured at discrete time sub-interval of t, call this unit of time d, which might be each day, with potentially t representing month and y possibly representing year.
    - Define $d \in \{1, \ldots, D_y\}$ as the index for discretized τ.
    - $z_m(\tau)$: Environmental exposure at moment in time τ.
    - $z_{m,y,t,d}$: Environmental exposure measured at location m, during period y, movement/mobility sub-period t, and environmental exposure measurement sub-sub-period d.

Second, we now discussion three cases or scenarios for the relationships between y, t, and d:

- **Case 1**: $T_y = 1$ and $D_y > 1$
    - e.g: 5-year or 10-year census
    - This means that we do not have locational movement variations for individuals in the current dataset, but we do have multiple measurements of environmental exposures within each period y we are interested in.
    - This might be the case if we have some administrative (census) type data measured in some year y only once, so we know the demographic distribution across locations in that year y, and additionally, we have daily ($D_y = 365$) or monthly ($D_y = 12$) observed environmental data across the M locations reporting population distributions.
- **Case 2**: $T_y > 1$ and $D_y = 1$
    - e.g.: high-frequency panel and low-frequency environmental data
    - This means that we do have locational movement data available, but only observe environmental exposures once during each period of potential movement.
    - This might be the case if we have panel data at the monthly level over a year, $T_y = 12$, and our pollution measure is observed once a month. Perhaps there is underlying data at higher frequency, but due to data restrictions, we are only given the average exposure at the monthly level.
- **Case 3**: $T_y > 1$ and $D_y > 1$:
    - e.g.: panel data + high-frequency environmental data
    - This means that we have location movement data available, and we also have multiple environmental exposures observed with each period of potential movement.
    - This might be the case if we have panel data at the annual level over several years (so the y period of time spans several years perhaps), and our pollution measures is at monthly or daily levels.

1.2.4 Exposure Data Input Matrixes

Have discretized the frequency of environmental data measurements, we can now consider storing the data in matrixes and computing locational-path individual-specific exposure moments based on matrixes of observed environmental exposure data.

We build two matrixes here. In the first matrix, there are M rows for the M locations, and there are $(T_y \cdot D_y)$ number of rows for all feasible environmental measurements:

- $\mathbf{Z}_y^{\text{loc}}$ matrix:
    - $\mathbf{Z}_y^{\text{loc}}$ is a $(M \times (T_y \cdot D_y))$ matrix with M rows and $(T_y \cdot D_y)$ columns.

– In each row m, we store environmental data over time dimensions d and t,

$$\left\{\{z_{m,y,t,d}\}_{d=1}^{D_y}\right\}_{t=1}^{T_y},$$

across the columns, with $(D_y \cdot (t-1)+d)$ as the column index.
– In **Case 1** above, the matrix would be M by $(1 \cdot D_y))$ in dimensions.

We also have an additional matrix, this matrix is not a data input matrix, but a matrix that we will generate by combining the information in the $\mathbf{Z}_y^{\text{loc}}$ matrix we just discussed with the information in the locational-path $\mathbf{Q}_y$ matrix we discussed earlier.

The second matrix is $\mathbf{Z}_y^{\text{path}}$.

- $\mathbf{Z}_y^{\text{path}}$ matrix:
  – $\mathbf{Z}_y^{\text{path}}$ is $(O \times (T_y \cdot D_y))$ matrix with O rows and $(T_y \cdot D_y)$ columns, note this has O rows based on the number of locational-paths.
  – At each row o, for column $(d \cdot (t-1)+d)$ through column $(d \cdot t)$, we copy over the $\{z_{m,y,t,d}\}_{d=1}^{D_y}$ data from matrix $\mathbf{Z}_y^{\text{loc}}$'s $m_{y,t}(o)^{th}$ row and $(D_y \cdot (t-1)+1)^{th}$ through $(D_y \cdot t)^{th}$ columns. Note that $m_{y,t}(o)$ is the location information for path o at time t from the $\mathbf{Q}_y$ matrix.
  – In each row o, we store environmental data from different locations, based on information in $\mathbf{Q}_y$. This means, for row o:

$$\left\{\left\{\left(\sum_{\widehat{m}=1}^{M} z_{\widehat{m},y,t,d} \cdot \mathbf{1}_{\{m_{t,y}(o)=\widehat{m}\}}\right)\right\}_{d=1}^{D_y}\right\}_{t=1}^{T_y}$$

is stored across the columns, with $(D_y \cdot (t-1)+d)$ as the column index.
  – $\mathbf{Z}_y^{\text{path}}$ can be generated based on $\mathbf{Z}_y^{\text{loc}}$ and $\mathbf{Q}_y$ after several wide to long, merge, and long to wide operations in standard database management programs.

Note that when individuals do not move across locations within period y, we have the simple case that $\mathbf{Z}_y^{\text{path}} = \mathbf{Z}_y^{\text{loc}}$, meaning the two matrixes are identical with $O = M$ as described earlier. This would be the situation in Case 1.

1.3 Simulating Exposure Distributions

Having defined key variables, distributions, and distributional matrixes, we can now simulate the structures just described to generate locational-path individual-specific moments of environmental exposures.

1.3.1 Case 1, $T_y = 1$ and $D_y > 1$

Now, we simulating the exposure data structure for Case 1: $T_y = 1$ and $D_y = 12$. For this situation, since there is no location movement information, as discussed prior, $M = O$ by construction. This can be the case if we have census data from some year, and environmental measurements over 12 months in that time-frame.

We simulate monthly exposure measurements across all locations. We will allow for different means and standard deviations of pollution distribution across locations, and we will also allow for multiple peaks via allowing for location-specific mixture distributions. Specifically, the strategy is:

1. For each location m, we will draw $(12 \cdot \text{MixCnt})$ random numbers from some location-specific log normal distribution.
2. We draw MixCnt discrete random integers uniformly between 1 and 11. Each number will be used as the *peak* point for each one of the MixCnt sets of 12 random numbers.
3. Suppose MixCnt = 2, and we drew 3 and 7 along with 24 random log-normally distributed numbers.
   - For the first set of 12 numbers, we sort the first 3 numbers in ascending order, and we sort the 4th through the 12th randomly drawn numbers in descending order, this creates a single peaked sequence of exposures with peak at the 3rd or 4th number/month.

- For the second set of 12 numbers, we sort the first 7 numbers in ascending order, and we sort the 8th through the 12th randomly drawn numbers in descending order, this creates a second single peaked shape with peak at the 7th or 8th number.
- For the distribution of environmental exposures for this location m across months, we take the weighted average month by month of the two single-peaked sets of 12 numbers we just generated.

Following the procedure described, the resulting exposure draws might have peak in the 2nd and 10th month, might have a single peak close to the middle or the year, might be very flat looking, etc. This accounts for the potentially diverse patterns of environmental exposure distributions across time within each location m.

Now we implement the procedure just described. First, define M and D_y.

```
# Locations and Time periods
it_M_location <- 10
it_D_y <- 12
# Locations to display
it_loc_disp <- 10
ar_it_loc_disp <- seq(1, it_M_location, length.out=it_loc_disp)
ar_it_loc_disp <- round(ar_it_loc_disp)
# Number of time periods to display
it_time_disp <- 12
ar_it_time_disp <- seq(1, it_D_y, length.out=it_time_disp)
ar_it_time_disp <- round(ar_it_time_disp)
```

Second, generate location specific log-normal means and standard deviations.

```
# Define parameters
fl_mean_meanlog_acrossM <- 3.4
fl_sd_meanlog_acrossM <- 0.5
fl_mean_sdlog_acrossM <- 0.5
fl_sd_sdlog_acrossM <- 0.1
# Vector of means and sds by locations
set.seed(123)
# random means
ar_meanlog_acrossM <- log(rlnorm(
  it_M_location, meanlog = fl_mean_meanlog_acrossM, sdlog=fl_sd_meanlog_acrossM))
# random sd
ar_sdlog_acrossM <- log(rlnorm(
  it_M_location, meanlog = fl_mean_sdlog_acrossM, sdlog=fl_sd_sdlog_acrossM))
# random peaks
it_mix_cnt <- 2
mt_it_peak <- matrix(sample(seq(1, it_D_y-1), it_M_location*it_mix_cnt, replace=TRUE),
                     nrow=it_M_location, ncol=it_mix_cnt)
# print
print(paste(round(ar_meanlog_acrossM,2)))
```

```
##  [1] "3.12" "3.28" "4.18" "3.44" "3.46" "4.26" "3.63" "2.77" "3.06" "3.18"
```

```
print(paste(round(ar_sdlog_acrossM,2)))
```

```
##  [1] "0.62" "0.54" "0.54" "0.51" "0.44" "0.68" "0.55" "0.3"  "0.57" "0.45"
```

```
print(mt_it_peak)
```

```
##       [,1] [,2]
##  [1,]   10   11
##  [2,]    7    6
##  [3,]    9    9
```

```
##  [4,]    9    2
##  [5,]   10    5
##  [6,]    7    8
##  [7,]   11    2
##  [8,]    5    1
##  [9,]    7    9
## [10,]    5   11
```

Third, we generate the $\mathbf{Z}_y = \mathbf{Z}_y^{\text{path}} = \mathbf{Z}_y^{\text{loc}}$ matrix. Note that they are the same since we are in Case 1. We are now following the strategies described earlier in this section.

```
# 1. Generate matrix to be filled
mt_Z_cone <- matrix(data=NA, nrow=it_M_location, ncol=it_D_y)
rownames(mt_Z_cone) <- paste0('m=', seq(1,it_M_location))
colnames(mt_Z_cone) <- paste0('d=', seq(1,it_D_y))

# 2. Fill matrix row by row
set.seed(456)
for (it_m in seq(1, it_M_location)){
  # Get mean and sd
  fl_meanlog <- ar_meanlog_acrossM[it_m]
  fl_sdlog <- ar_sdlog_acrossM[it_m]
  # Get peaks
  ar_it_peaks <- mt_it_peak[it_m,]
  # Generate random number, 2 x it_D_y for mixture of it_mix_cnt distributions
  mt_fl_draws <- matrix(rlnorm(it_D_y*it_mix_cnt, meanlog=fl_meanlog, sdlog=fl_sdlog),
                        nrow=it_mix_cnt, ncol=it_D_y)
  # Resort ascending before peak, descending after peak
  for (it_mix_ctr in seq(1,it_mix_cnt)){
    it_peak <- ar_it_peaks[it_mix_ctr]
    mt_fl_draws[it_mix_ctr, 1:it_peak] <- sort(mt_fl_draws[it_mix_ctr, 1:it_peak])
    mt_fl_draws[it_mix_ctr, seq(it_peak+1, it_D_y)] <- sort(
      mt_fl_draws[it_mix_ctr, seq(it_peak+1, it_D_y)],
      decreasing=TRUE)
  }
  # Average across mixtures, equal weights
  ar_fl_draws_mix_weighted <- colSums(mt_fl_draws)/it_mix_cnt
  # Fill matrix
  mt_Z_cone[it_m, ] <- ar_fl_draws_mix_weighted
}

# 3. tibble and display
# Combine to tibble, add name col1, col2, etc.
ar_st_varnames <- c('location', colnames(mt_Z_cone))
tb_Z_cone <- as_tibble(mt_Z_cone) %>%
  rowid_to_column(var = "id") %>%
  rename_all(~c(ar_st_varnames)) %>%
  mutate(location = paste0('m=', location))

# Display
st_caption = "PM10 exposure across locations and time"
tb_Z_cone[ar_it_loc_disp,c(1,ar_it_time_disp+1)] %>%
  kable(caption = st_caption) %>% kable_styling_fc_wide()
```

PM10 exposure across locations and time

location	d=1	d=2	d=3	d=4	d=5	d=6	d=7	d=8	d=9	d=10	d=11	d=12
m=1	8.504895	9.67479	15.65469	20.48814	31.80818	34.80976	37.61727	46.79303	62.85414	78.48371	46.46616	17.544852
m=2	12.021470	19.08505	21.25595	26.06994	37.62492	49.96870	49.76649	57.17538	28.47315	27.11189	24.76007	23.025230
m=3	28.882246	45.74821	63.54363	70.73737	77.79342	84.89134	96.77687	121.37000	126.63471	107.93362	51.33159	39.415826
m=4	24.259793	39.72569	40.88998	33.99299	38.61860	32.74619	32.85029	37.18869	57.33850	43.46071	23.20980	20.698045
m=5	22.378126	25.24594	29.57953	34.67750	39.81435	42.67476	41.58195	41.61536	39.63658	41.09408	35.91566	18.249551
m=6	40.104662	45.70796	52.94881	62.73639	66.60968	106.87865	125.92674	166.57880	108.69790	50.88886	45.98992	35.103694
m=7	17.857811	38.48846	38.27401	35.77649	35.89942	36.48475	44.68330	47.76937	47.25462	46.98947	53.66579	57.971892
m=8	12.268494	15.87203	15.95402	17.03461	21.92676	21.26708	19.75517	16.48595	14.67311	13.64343	11.83509	9.825761
m=9	13.923821	15.30081	18.23575	19.02678	23.83729	28.03411	37.58018	40.29858	37.42698	34.54124	26.15540	15.451811
m=10	11.581023	17.35950	18.69695	25.11728	33.19108	46.20657	41.26486	38.30603	34.16592	32.72968	33.62844	15.638355

1.4 Within Individual Exposure Moments

We now compute within-individual exposure moments. The computation of these moments are based on the $\mathbf{Z}_y^{\text{path}}$ matrix, which again is the same as the $\mathbf{Z}_y^{\text{loc}}$ matrix for Case 1. Regardless of which case we are looking at, we compute these statistics in the same way.

Below, we compute row-specific distributional statistics. We are interested in the mean and percentiles. Specifically, we want to compute:

1. mean: which is simply the average over columns for each row
2. p20: the 20th percentile
3. p50: the 50th percentile
4. p80: the 80th percentile

We compute percentiles based on the nearest-rank method. When there are 12 months of data, this means this 20th percentile is based on the exposure measure in the month with the 3rd lowest exposure (1st is the lowest), the 50th percentile is based on the exposure measure in the month with the 6th lowest exposure, and the 80th percentile is based on the exposure measure in the month with the 3rd highest exposure.

In the code below, we selected the percentiles we are interested in, and also compute the mean. We construct a matrix that stores for each locational-path (or location in Case 1) row, four statistics across columns for the locational-path within-individual specific exposure moments.

```
# 1. number of quantiles of interest
ar_quantiles <- c(0.2, 0.5, 0.8)
it_quantiles <- length(ar_quantiles)

# 2. Generate matrix to be filled
mt_S_moments <- matrix(data=NA, nrow=it_M_location, ncol=1+it_quantiles)
rownames(mt_S_moments) <- paste0('m=', seq(1,it_M_location))
colnames(mt_S_moments) <- c('pm_indi_mean', paste0('pm_indi_q', round(ar_quantiles*100,0)))

# 3. Compute quantiles
for (it_m in seq(1, it_M_location)){
  ar_Z <- mt_Z_cone[it_m, ]
  fl_mean <- mean(ar_Z, na.rm=TRUE)
  # note we use type=1, this uses the nearest-rank method
  ar_quant_vals <- stats::quantile(ar_Z, probs=ar_quantiles, na.rm = TRUE, type=1)
  mt_S_moments[it_m,] <- c(fl_mean, ar_quant_vals)
}
```

Now we display results. The table shown here, corresponds to the "PM10 Exposure across locations (10 of 20 locations shown)" table from the Simulate Environmental Exposure section on the Share of Environmental Exposure Burden Across Population Groups page.

```
# Column Names
ar_st_varnames <- c("locational_path", colnames(mt_S_moments))

# Combine to tibble, add name col1, col2, etc.
tb_loc_indi_dist <- as_tibble(mt_S_moments) %>%
  rowid_to_column(var = "id") %>%
  rename_all(~c(ar_st_varnames)) %>%
  mutate(locational_path = paste0("locational_path=", locational_path))

# Display
st_caption = "PM10 exposure individual-moments across locations/paths"
tb_loc_indi_dist[ar_it_loc_disp,] %>%
  kable(caption = st_caption) %>% kable_styling_fc_wide()
```

PM10 exposure individual-moments across locations/paths

locational_path	pm_indi_mean	pm_indi_q20	pm_indi_q50	pm_indi_q80
locational_path=1	34.22497	15.65469	31.80818	46.79303
locational_path=2	31.36152	21.25595	26.06994	49.76649
locational_path=3	76.25490	45.74821	70.73737	107.93362
locational_path=4	35.41494	24.25979	33.99299	40.88998
locational_path=5	34.37195	25.24594	35.91566	41.58195
locational_path=6	75.68101	45.70796	52.94881	108.69790
locational_path=7	41.75962	35.89942	38.48846	47.76937
locational_path=8	15.87846	12.26849	15.87203	19.75517
locational_path=9	25.81773	15.45181	23.83729	37.42698
locational_path=10	28.99047	17.35950	32.72968	38.30603

In the "PM10 Exposure across locations (10 of 20 locations shown)" table, we had only two columns, the first is for the M locations, and the second column is just the mean measurement for each location (we did not consider sub-periods of time there, so it was just a single measurement for each location). Our results here generalizes the analysis on the Share of Environmental Exposure Burden Across Population Groups page, which does not consider within-individual or locational-path distributions. All the analysis there are based on distributions of individual mean exposures. Our analysis here is more general because:

1. We now also consider other moments of individual-specific distributions.
   - We can grab out any of the moment columns from the table below, replace the "avgdailypm10" column in "PM10 Exposure across locations (10 of 20 locations shown)" table with the new statistics, and proceed with the same analysis from page Share of Environmental Exposure Burden Across Population Groups.
   - Two locations/individuals with the same mean pm2.5 within a year might have very different p20, p50, and p80. One location could have identical pollution each month, the other location could have substantially varying pollution across months.
   - If what matters for health is extreme exposure, rather than cumulative exposure over time period y, then within-individual (or locational-path) distributions matters.
2. We now also consider movement/mobility within period y:
   - Our structure works when individuals do not move and $\mathbf{Z}_y = \mathbf{Z}_y^{\text{path}} = \mathbf{Z}_y^{\text{loc}}$, this means the prior data structure is accommodated
   - Our framework also works when survey panel data shows individuals moving within time period y. This is generalization, a weakening of the input data requirement for our analysis.
   - It turns out to accommodate movements, we do not need to modify any of the computations on page Share of Environmental Exposure Burden Across Population Groups, we just need to reinterpret

the M location-specific rows there for the input data matrixes as the O locational-path-specific rows. The code there does not care about what interpretations the rows have.

1.5 Within Individual Adjusted Mean

Previously we computed moments of the individual distribution. Given heterogeneity in the distribution, we could compute alternative means where exposures in months below some exposure threshold is set to 0. Note that this is not computing a conditional expectation.

We take the "PM10 exposure across locations and time" matrix we computed from prior as given.

```
# 1. number of quantiles of interest
ar_thresholds <- seq(0, 90, by=10)
ar_thresholds <- c(0, 15, 35, 50, 75)
it_thresholds <- length(ar_thresholds)

# 2. Generate matrix to be filled
mt_S_mean_thres <- matrix(data=NA, nrow=it_M_location, ncol=it_thresholds)
rownames(mt_S_mean_thres) <- paste0('m=', seq(1,it_M_location))
colnames(mt_S_mean_thres) <- paste0('pm_indi_thr', ar_thresholds)

# 3. Compute quantiles
for (it_m in seq(1, it_M_location)){
  ar_Z <- mt_Z_cone[it_m, ]
  fl_mean <- mean(ar_Z)
  ar_mean_thres <- c()
  for (it_thres in seq(1, it_thresholds)){
    ar_Z_thres <- ar_Z
    ar_Z_thres[ar_Z < ar_thresholds[it_thres]] <- 0
    fl_mean_thres <- mean(ar_Z_thres, na.rm = T)
    ar_mean_thres <- c(ar_mean_thres, fl_mean_thres)
  }
  # note we use type=1, this uses the nearest-rank method
  mt_S_mean_thres[it_m,] <- ar_mean_thres
}
```

Now we display results. Similar to the moments based results, the table shown here corresponds to the "PM10 Exposure across locations (10 of 20 locations shown)" table from the Simulate Environmental Exposure section on the Share of Environmental Exposure Burden Across Population Groups page.

```
# Column Names
ar_st_varnames <- c("locational_path", colnames(mt_S_mean_thres))

# Combine to tibble, add name col1, col2, etc.
tb_loc_indi_dist_thres <- as_tibble(mt_S_mean_thres) %>%
  rowid_to_column(var = "id") %>%
  rename_all(~c(ar_st_varnames)) %>%
  mutate(locational_path = paste0("locational_path=", locational_path))

# Display
st_caption = "PM10 exposure individual < threshold to 0 means across locations/paths"
tb_loc_indi_dist_thres[ar_it_loc_disp,] %>%
  kable(caption = st_caption) %>% kable_styling_fc_wide()
```

PM10 exposure individual < threshold to 0 means across locations/paths

locational_path	pm_indi_thr0	pm_indi_thr15	pm_indi_thr35	pm_indi_thr50	pm_indi_thr75
locational_path=1	34.22497	32.70999	22.684526	11.778154	6.540309
locational_path=2	31.36152	30.35973	16.211290	4.764615	0.000000
locational_path=3	76.25490	76.25490	73.848049	66.751046	51.283330
locational_path=4	35.41494	35.41494	21.435181	4.778208	0.000000
locational_path=5	34.37195	34.37195	23.527728	0.000000	0.000000
locational_path=6	75.68101	75.68101	75.681007	61.772154	42.340175
locational_path=7	41.75962	41.75962	40.271465	9.303140	0.000000
locational_path=8	15.87846	10.69130	0.000000	0.000000	0.000000
locational_path=9	25.81773	24.65741	9.608812	0.000000	0.000000
locational_path=10	28.99047	28.02539	10.481454	0.000000	0.000000