Analysis on the Effect of Meteorological Factors on Ozone Concentration Using Varying-Coefficient Quantile Regression

Matteo Venturini

June 8, 2025

Student Number: 2469579

Professor: Yudhie Andriyana

Course: Non-parametric methods

Hasselt University

1 Introduction and Data Description

One way to analyze the relationship between variables is by using regression analysis. Parametric regression is mostly used when the relationship between the regressor and the outcome variable is linear, and one of the most important assumptions is that the conditional expectation of Y given X, $E[Y \mid X]$, follows a specified probability distribution with a finite number of parameters (Yavuz & Şahin, 2022). On the other hand, non-parametric techniques estimate the functional form directly from the data (Imam et al., 2024).

An interesting form on non-parametric regression is quantile regression, which instead of modeling the expected mean of the outcome variable, it can model the median and other quantiles of the distribution, allowing for a more accurate representation of the effect of the regressor while also being more robust against outliers (Petscher & Logan, 2013).

While standard quantile regression models different parts of an outcome's distribution, sometimes the quantile curves cross each other, which is theoretically inconsistent. The work by Andriyana, Gijbels, and Verhasselt (2014) addresses this within flexible varying-coefficient models, developing methods to ensure non-crossing curves. Their paper further extends these models to account for heteroscedasticity, where the error variability V(T) can change over time, allowing for a more complete understanding of the conditional distribution.

This approach, which uses P-splines for estimation, is particularly suited for analyzing dynamic relationships, such as those explored in the airquality dataset where Ozone levels and their variability are influenced by changing meteorological factors over the study period.

The model equation used in this analysis is:

$$Y_{\text{Ozone}}(t) = \beta_0(t) + \beta_1(t)X_{\text{Solar.R}}(t) + \beta_2(t)X_{\text{Wind}}(t) + \beta_3(t)X_{\text{Temp}}(t) + V(t)\varepsilon(t)$$

The ‘airquality‘ dataset from R is used. It contains 153 daily observations of air quality measurements in New York from May to September 1973. After removing rows with missing values for Ozone, Solar.R, Wind, and Temp, 111 complete observations remain. The response variable is Ozone concentration. The time variable (‘times‘) is the day of the study (1 to 111). Covariates include Solar Radiation (‘Solar.R‘), Wind Speed (‘Wind‘), and Temperature (‘Temp‘). The ‘AHeVT‘ function from the ‘QRegVCM‘ package is employed for the analysis. The first measurements at day 1 were made on the 1st of May; measurements at day 60 were made on the 29th of June; measurements at day 111 were made on the 19th of August.

2 Data Analysis

As a preliminary step, the distributions of the dependent and independent variables were examined using non-parametric kernel density estimation. Following this, the bivariate relationships between the outcome and each predictor were modeled using both Nadaraya-Watson (NWE) and Local Polynomial (LP) regression. Lastly, the Varying-Coefficient Quantile Regression model was tested using the 'AHeVT' function from the 'QRegVCM' package. The analysis was conducted using R version 4.5.0.

2.1 Kernel Density Estimation

Figure 1 displays the densities of the dependent and independent variables. We can clearly see that a parametric distribution would not fit all our data well: the densities of Ozone and Wind are right-skewed, while Solar Radiation shows a broad, bimodal distribution. This suggests that using kernel density estimation was an appropriate exploratory approach.

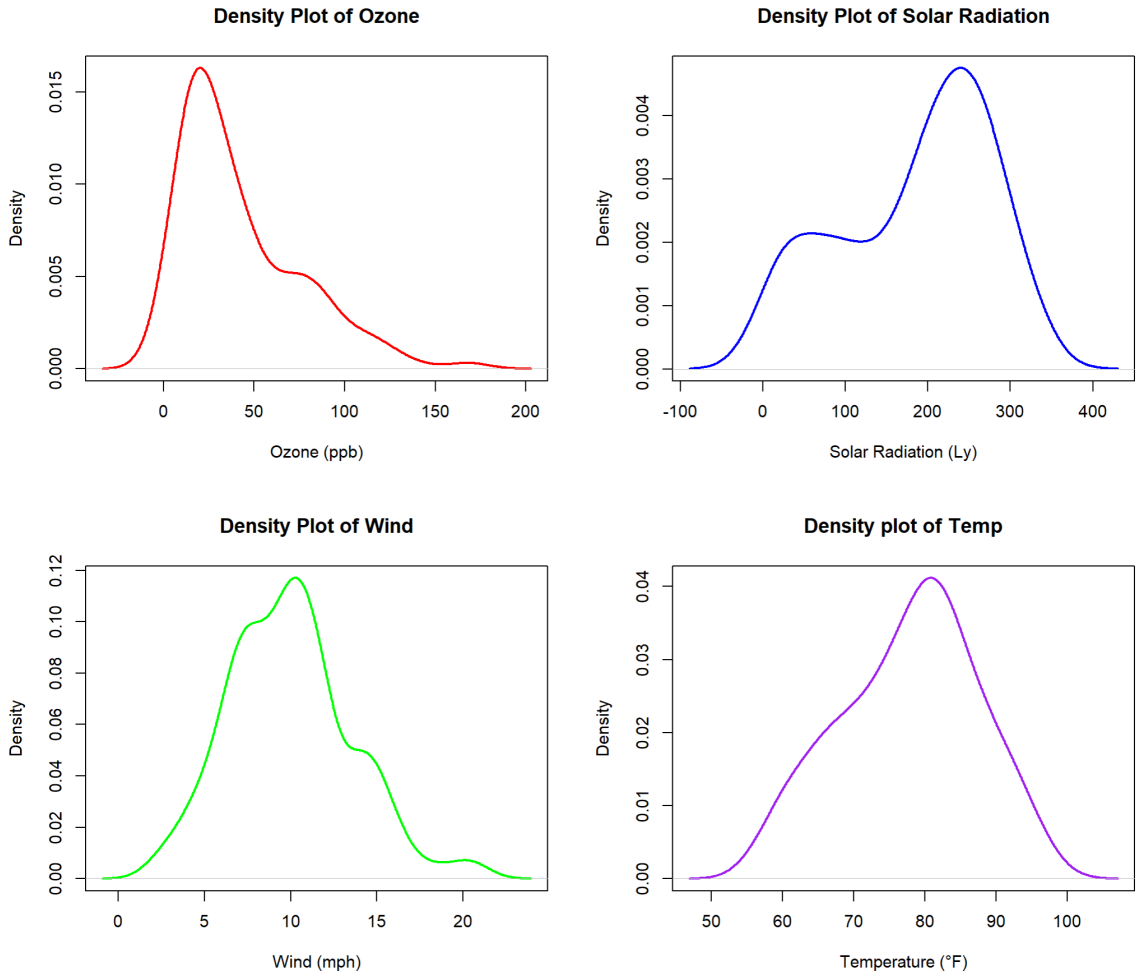


Figure 1: Densities of Ozone (DV), Solar Radiation (IV), Wind (IV), and Temperature (IV) estimated using Kernel Density Estimation

2.2 Ozone-Covariate non-parametric regression

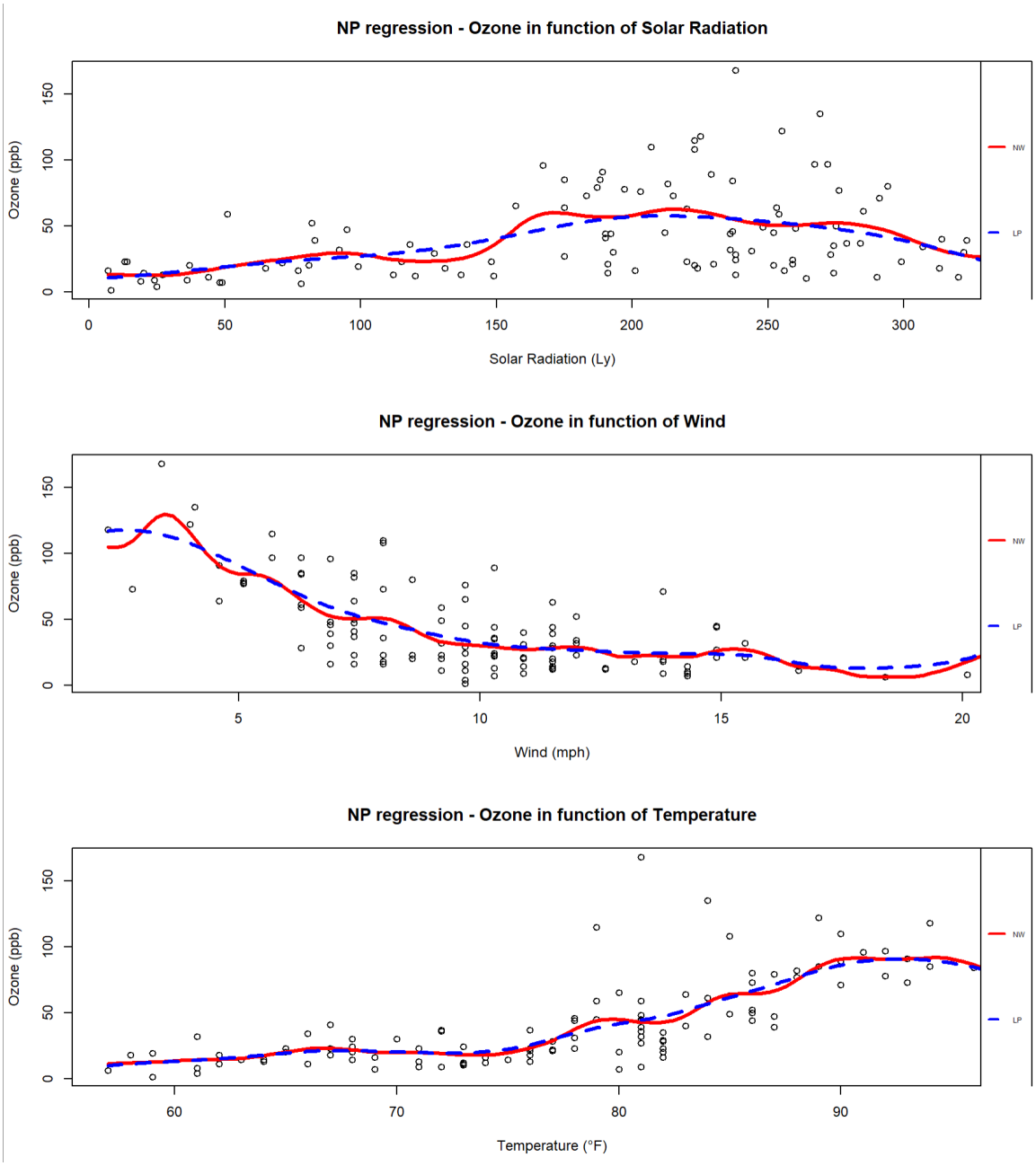
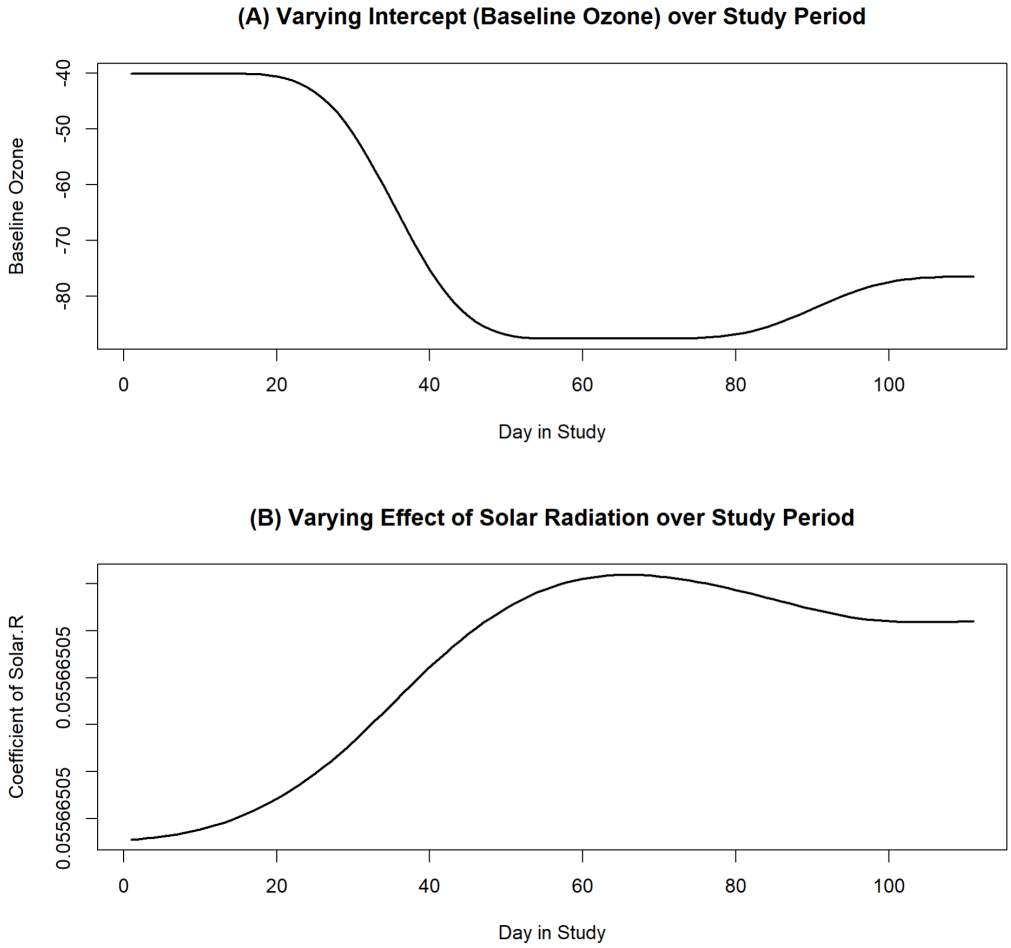


Figure 2: Non-parametric regression of Ozone with each covariate using Nadaraya-Watson (NW) and Local Polynomial (LP) regression

Before fitting the full varying-coefficient model, the individual relationships between Ozone concentration and each meteorological predictor were explored using nonparametric regression techniques. Specifically, Nadaraya-Watson and Local Polynomial regression fits were applied to visualize these associations (Figure 2). For Solar Radiation, the NW (red) and LP (blue) fits suggest that Ozone generally increases with solar radiation up to approximately 200-250 Ly, beyond which the relationship appears to plateau and then slightly decline. Wind speed shows a clear negative association, with Ozone levels decreasing most sharply as wind increases from low values, and this effect lessens at higher

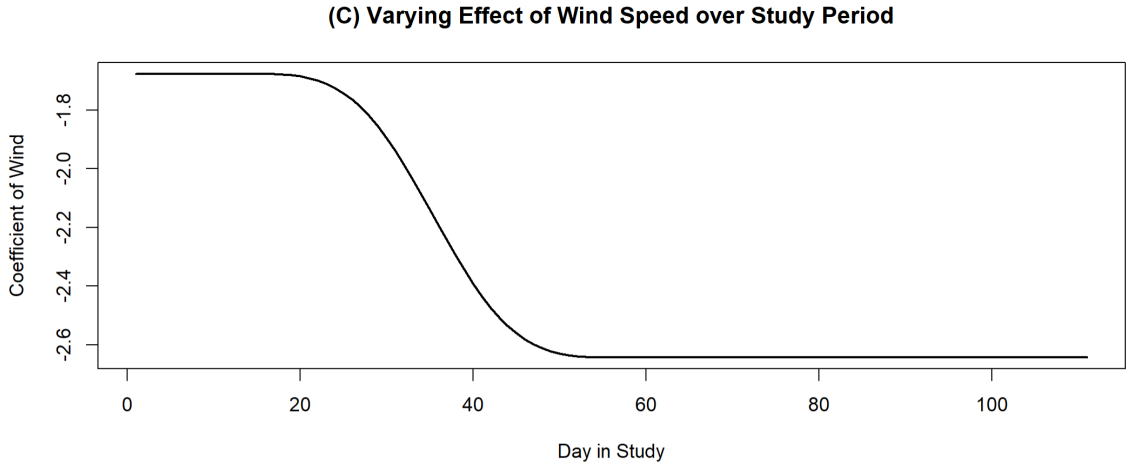
wind speeds. Temperature exhibits a consistently positive relationship with Ozone, with the increase in Ozone becoming more pronounced at temperatures above 75-80°F. These initial visualizations highlight the non-linear nature of the relationships between Ozone and the individual predictors, motivating the use of a flexible varying-coefficient modeling approach.

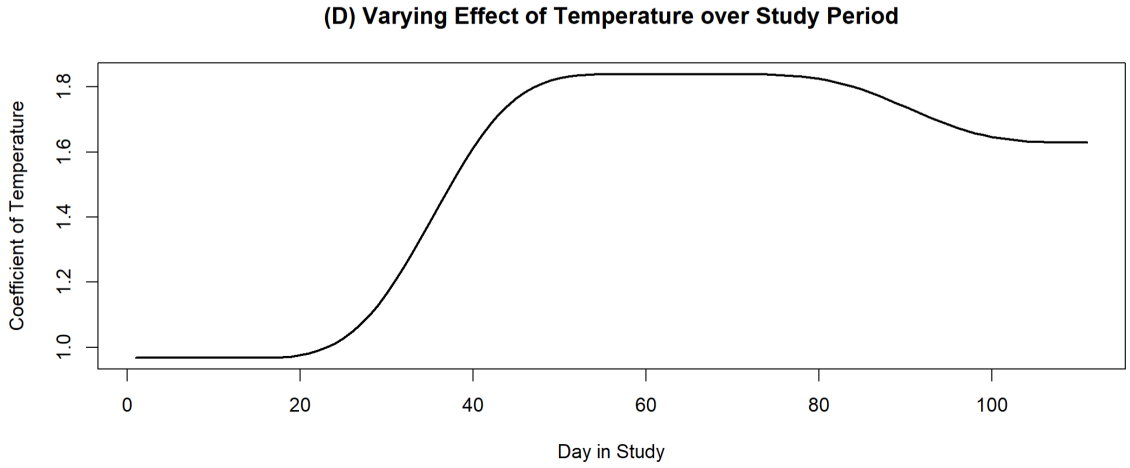
2.3 Varying Coefficients



The results from the primary Varying-Coefficient Quantile Regression model (AHeVT) are now presented. Plot A displays the baseline Ozone intercept. Given the scale and the primary focus on covariate effects, this intercept will not be extensively discussed. More interesting is plot B, which shows the varying effect of Solar Radiation on Ozone levels as a function of day.

We observe that throughout the study period, the effect of Solar Radiation has consistently been positive, meaning that higher levels of Solar Radiation are associated with increased Ozone concentrations. Notably, this effect is not constant over time but increases during the summer months, peaking in July. However, it is worth mentioning that the magnitude of the coefficient remains quite small, indicating an overall limited effect. These results are in line with what was observed in Figure 2 using NWE and LP regression.





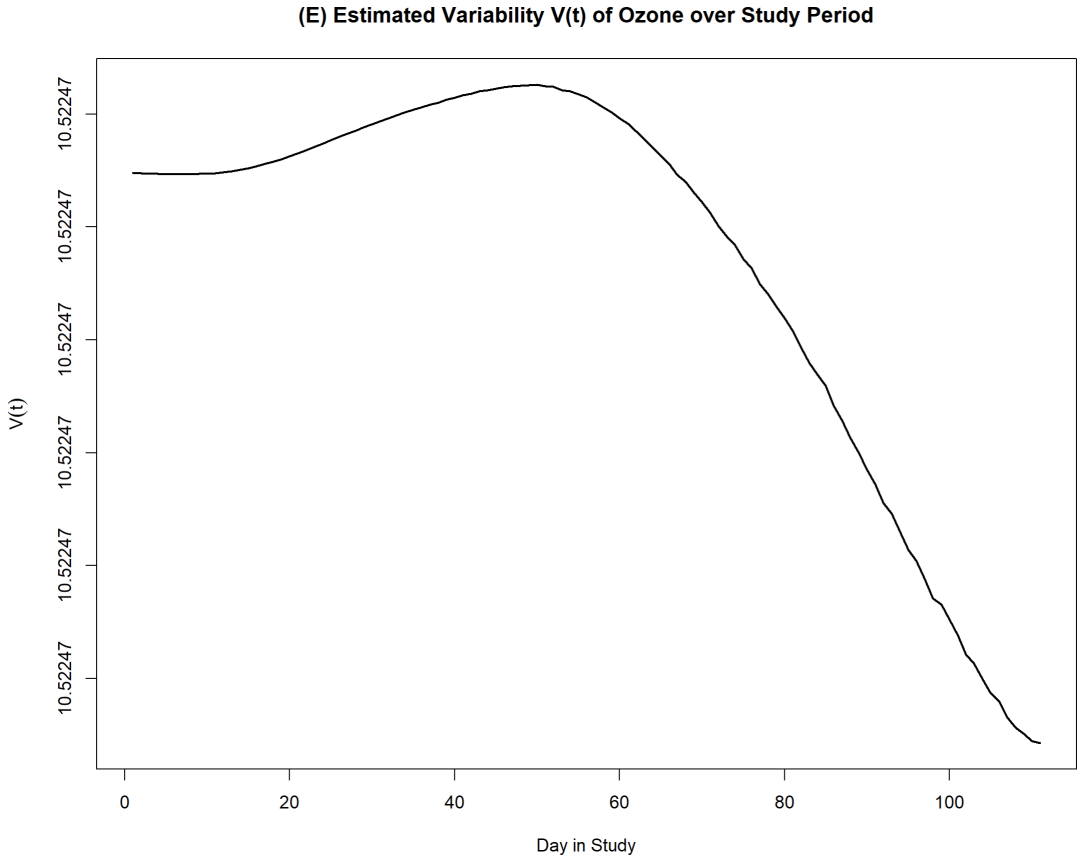
In plot C, we can observe the varying effect of wind speed on ozone levels. The coefficient is negative, indicating that stronger winds reduce ozone concentrations. The magnitude of this coefficient is relatively large compared to the other regressors. It is also evident that, with the onset of summer, the effect of wind becomes more pronounced until it eventually plateaus. Also this result is in line with what was observed in Figure 2.

Plot D shows the varying effect of temperature. Here as well, the coefficient is positive, meaning that higher temperatures are associated with increased ozone levels. As summer begins, the coefficient rises, then plateaus, and eventually declines around August, which aligns with what was observed using NWE and LP regression.

2.4 Estimated Variability V(t)

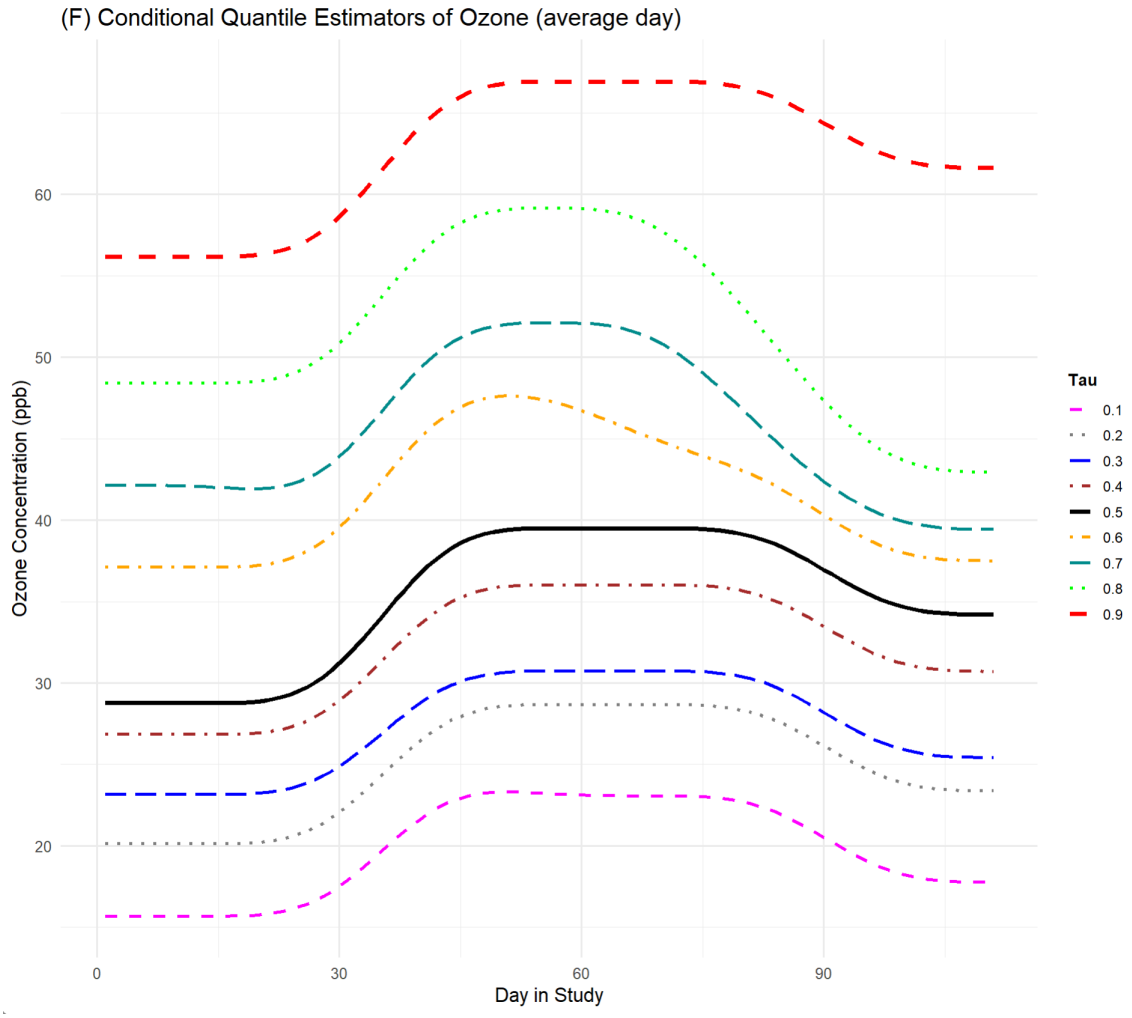
The estimated variability function, $V(t)$, indicates how the overall spread of the Ozone concentration distribution changes over the study period. In plot E it is possible to see how does the magnitude of $V(t)$ change as function of day. We can see that during Spring it starts high, it peaks during early Summer, and then it starts to fall during July. Despite this relative change, the absolute change in magnitude of $V(t)$ is minimal. This means that despite the model can handle heteroskedasticity, the data suggests that the error term is being approximately multiplied by a constant rather than a variable. Fitting a

general hetersocedastic varying-coefficient model might yield different results.

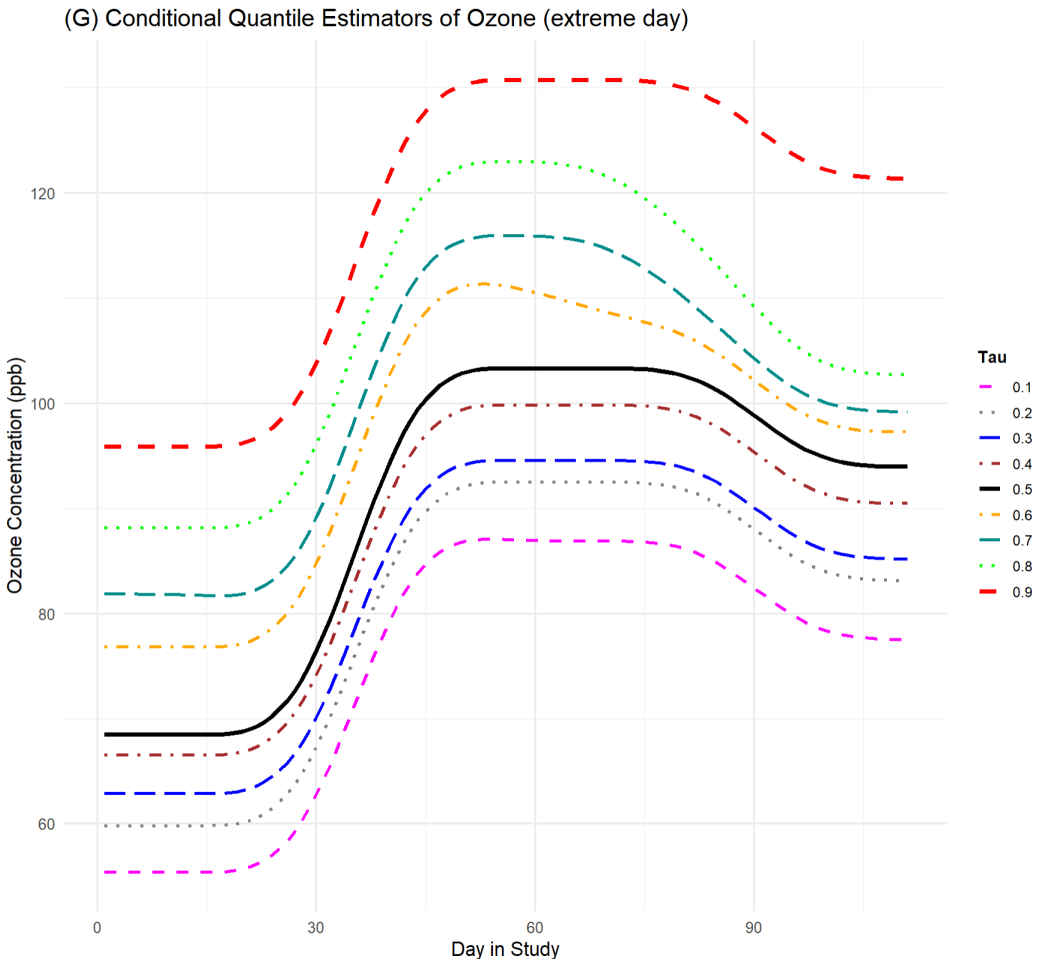


2.5 Conditional Quantile Estimators

The conditional quantile estimators show how different percentiles of the Ozone distribution evolve over the study period, given average levels of solar radiation, wind speed, and temperature. From plot F it is possible to see that Ozone concentration exhibit a strong seasonal pattern, peaking during early Summer. The variability around the median value is pretty constant, as the lines are more or less equally spaced throughout the study period. Given that the lines do not cross each other, it is possible to have a detailed look at how different parts of the Ozone distribution respond over the season for this specific average-day scenario.



In plot G we can see instead the conditional quantile estimators given the worst case scenario, with the highest value of Solar Radiation, lowest value of Wind, and highest value of Temperature. While it is unlikely that all these conditions occur at once, it is relevant to take into account its possibility.


(G) Conditional Quantile Estimators of Ozone (extreme day)
Ozone Concentration (ppb)
120
100
80
60
0
30
60
90
Day in Study
Tau
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9

3 Conclusions

The analysis of the ‘airquality‘ dataset was initiated with an exploration of variable distributions and bivariate relationships, before delving into the application of varying-coefficient quantile regression which revealed dynamic seasonal patterns in Ozone concentration and the influence of meteorological factors. The initial kernel density estimations highlighted the non-normal characteristics of several variables, particularly the skewed distributions of Ozone and Wind, and the bimodal distribution of Solar Radiation, underscoring the suitability of non-parametric approaches. Furthermore, preliminary Nadaraya-Watson and Local Polynomial regressions visually confirmed the non-linear relationships between Ozone and individual meteorological predictors, providing a foundation for the more complex VCM analysis. The VCM analysis showed that median Ozone concentrations, even under average meteorological conditions, exhibit a strong seasonal trend, peaking during the summer months (around June-July) and declining in spring and fall. This underlying seasonality is very important when considering the risk of exposure. Solar Radiation consistently shows a positive, yet small, effect on increasing Ozone levels, with this influence being most pronounced during summer. Wind Speed has a notable negative effect on Ozone, and this effect becomes more accentuated from the onset of summer. Temperature also positively influences Ozone levels, with its impact strengthening during summer before slightly decreasing towards August. These varying coefficient patterns were broadly consistent with the initial trends observed in the bivariate nonparametric regressions. The estimated time-dependent variability component, $V(t)$, showed a peak in early summer followed by a decline. However, the absolute change in the magnitude of $V(t)$ was minimal, suggesting that the purely time-dependent aspect of heteroscedasticity in this dataset is limited. For days with average meteorological conditions, the entire conditional distribution of Ozone (from the 10th to 90th percentiles) shifts seasonally, peaking in summer. The spread between these quantiles, representing the range of likely Ozone values, also appeared relatively consistent throughout the study period under these average conditions, aligning with the near-constant $V(t)$. When considering a worst-case scenario (maximum Solar Radiation, minimum Wind, maximum Temperature), the model estimates significantly higher Ozone levels across all quantiles, highlighting the amplified risk under such conditions.

These findings have direct relevance to public health. Ozone is a potent oxidant with significant adverse effects on the respiratory system, even at low ambient concentrations (Swanson et al., 2022). The World Health Organization (WHO) recommends an 8-hour maximum of 100 µg/m^3 (approximately 50 ppb) (WHO, 2022). The analysis shows that median Ozone levels under average conditions can approach this threshold during summer peaks, and under worst-case meteorological scenarios, even lower quantiles can greatly exceed this guideline, with the 90th percentile reaching substantially higher values.

Individuals engaging in prolonged outdoor activities during summer months, when the positive impacts of solar radiation and temperature on Ozone concentration are strongest, face the highest cumulative exposure and risk (EPA, 2025). The model's ability to show how different parts of the Ozone distribution shift seasonally, and under different covariate scenarios, underscores the importance of dynamic air quality monitoring and advisories, particularly during periods identified as high-risk by such analyses. While the specific $V(t)$ component of variability was found to be minor in this ‘AHeVT‘ application, the significant variation in the conditional mean driven by the varying coefficients $\beta_k(t)$ high-

lights the changing nature of Ozone pollution throughout the year. Further investigation with models allowing for more complex heteroscedastic structures (e.g., $V(X, t)$) could provide additional insights into Ozone predictability.

References

[1] Andriyana, Y., Gijbels, I., & Verhasselt, A. (2014). P-splines quantile regression estimation in varying coefficient models. *TEST, 23*(1), 153–194. https://doi.org/10.1007/s11749-013-0346-2

[2] Andriyana, Y., & Ibrahim, M. A. (2018). *QRegVCM: Quantile Regression in Varying-Coefficient Models* (Version 1.2) [R package]. https://CRAN.R-project.org/package=QRegVCM

[3] Malley, C. S., Henze, D. K., Johan C.I. Kuylenstierna, Vallack, H. W., Davila, Y., Anenberg, S. C., Turner, M. C., & Ashmore, M. R. (2017). Updated Global Estimates of Respiratory Mortality in Adults ≥30 Years of Age Attributable to Long-Term Ozone Exposure. *Environmental Health Perspectives, 125*(8). https://doi.org/10.1289/ehp1390

[4] Petscher, Y., & Logan, J. A. R. (2013). Quantile regression in the study of developmental sciences. *Child Development, 85*(3), 861–881. https://doi.org/10.1111/cdev.12190

[5] Swanson, T. J., Jamal, Z., & Chapman, J. (2022, November 15). *Ozone Toxicity*. Nih.gov; StatPearls Publishing. https://www.ncbi.nlm.nih.gov/books/NBK430751/

[6] U.S. Environmental Protection Agency. (2025, March 27). *Health effects of Ozone in the general population.* https://www.epa.gov/Ozone-pollution-and-your-patients-health/health-effects-Ozone-general-population

[7] World Health Organization. (2022, May 2). *Ambient (outdoor) air quality and health.* https://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health

[8] Yavuz, E., & Şahin, M. (2022). Investigation of Parametric, Non-Parametric and Semiparametric Methods in Regression Analysis. *Sakarya University Journal of Science, 26*(6), 1111–1116. https://doi.org/10.16984/saufenbilder.1147135

4 Appendix

4.1 R Code

```
rm(list=ls())

library(QRegVCM)


data(airquality)



# Remove rows with NAs
airquality_clean <- na.omit(airquality[, c("Ozone", "Solar.R", "
   Wind", "Temp", "Month", "Day")])

# Create a new continuous variable for day
airquality_clean$DayInStudy <- 1:nrow(airquality_clean)


# Density estimation for IVs and DV
par(mfrow=c(2,2))
x1 <- airquality_clean$Ozone
fx1 = density(x1, kernel="gaussian")
plot(fx1, col="red", lwd=2,  main="Density␣Plot␣of␣Ozone",
     xlab="Ozone␣(ppb)")

x2 <- airquality_clean$Solar.R
fx2 = density(x2, kernel="gaussian")
plot(fx2, col="blue", lwd=2,  main="Density␣Plot␣of␣Solar␣
   Radiation",
     xlab="Solar␣Radiation␣(Ly)")

x3 <- airquality_clean$Wind
fx3 = density(x3, kernel="gaussian")
plot(fx3, col="green", lwd=2, main="Density␣Plot␣of␣Wind",
     xlab = "Wind␣(mph)")

x4 <- airquality_clean$Temp
fx4 = density(x4, kernel="gaussian")
plot(fx4, col="purple", lwd=2, main="Density␣plot␣of␣Temp",
     xlab = "Temperature␣( F )")

par(mfrow=c(3,1)) # I recommend "shrinking" the x axis when
   clicking on "Zoom", otherwise it looks
# too stretched out.


# NW and local polynomial regression
```

```
library("KernSmooth")

# Ozone and Solar Radiation
h1 = sd(x2)*(length(x2))^(-1/5) ## Rule of thumb bandwidth for NW
h2 <- dpill(x2, x1) ## Plug-in method for LP

plot(x2,x1, main='NP regression - Ozone in function of Solar
   Radiation',
     xlab = 'Solar Radiation (Ly)', ylab='Ozone (ppb)')
lines(ksmooth(x2, x1, kernel="normal", bandwidth=h1), col="red",
   lwd=3, lty=1)
lines(locpoly(x2, x1, degree=1,  kernel="normal", bandwidth=h2,
   range.x=range(x2), binned=FALSE),
      col="blue", lwd=3, lty=2)
legend("topright", c("NW", "LP"),ncol=1, col=c("red","blue"),
       lwd=c(2,2), lty=c(1,2), cex=0.5)


# Ozone and Wind
h1 = sd(x3)*(length(x3))^(-1/5) ## Rule of thumb bandwidth for NW
h2 <- dpill(x3, x1) ## Plug-in method for LP

plot(x3,x1, main='NP regression - Ozone in function of Wind',
     xlab = 'Wind (mph)', ylab='Ozone (ppb)')
lines(ksmooth(x3, x1, kernel="normal", bandwidth=h1), col="red",
   lwd=3, lty=1)
lines(locpoly(x3, x1, degree=1,  kernel="normal", bandwidth=h2,
   range.x=range(x3), binned=FALSE),
      col="blue", lwd=3, lty=2)
legend("topright", c("NW", "LP"),ncol=1, col=c("red","blue"),
       lwd=c(2,2), lty=c(1,2), cex=0.5)

# Ozone and Temperature
h1 = sd(x4)*(length(x4))^(-1/5) ## Rule of thumb bandwidth for NW
h2 <- dpill(x4, x1) ## Plug-in method for LP

plot(x4,x1, main='NP regression - Ozone in function of
   Temperature',
     xlab = 'Temperature ( F )', ylab='Ozone (ppb)')
lines(ksmooth(x4, x1, kernel="normal", bandwidth=h1), col="red",
   lwd=3, lty=1)
lines(locpoly(x4, x1, degree=1,  kernel="normal", bandwidth=h2,
   range.x=range(x4), binned=FALSE),
      col="blue", lwd=3, lty=2)
legend("topright", c("NW", "LP"),ncol=1, col=c("red","blue"),
       lwd=c(2,2), lty=c(1,2), cex=0.5)



par(mfrow=c(1,1))
```

```r
# AHeVT model
y <- airquality_clean$Ozone           # Ozone concentration (
   response)
times <- airquality_clean$DayInStudy  # Day in the study (time
   variable)
subj <- airquality_clean$Month        # Month (subject/grouping
   identifier)
dim <- length(y)                      # Number of observations
   after cleaning

# Define covariates
x0 <- rep(1, dim)                       # Intercept
x1 <- airquality_clean$Solar.R          # Solar Radiation (
   continuous)
x2 <- airquality_clean$Wind             # Wind Speed (continuous)
x3 <- airquality_clean$Temp             # Temperature (continuous
   )
X <- cbind(x0, x1, x2, x3)

# Average day
#VecX <- c(1, mean(x1), mean(x2), mean(x3))

# Extreme day
VecX <- c(1, max(x1), min(x2), max(x3))


kn <- c(8, 8, 8, 8)           # Number of knots for each varying
   coefficient.
degree <- c(3, 3, 3, 3)       # Degree of B-spline basis
taus <- seq(0.1, 0.9, 0.1)    # Quantiles of interest
lambdas <- c(1, 1.5, 1.5, 1.5) # Smoothing parameters
d <- c(1, 1, 1, 1)            # Order of differencing operator
gam <- 1/2                    # Power used in estimating smoothing
    parameter


AHe_air <- AHeVT(VecX = VecX, times = times, subj = subj, X = X,
   y = y, d = d,
                 tau = taus, kn = kn, degree = degree, lambda =
                    lambdas, gam = gam)

hat_bt50_air <- AHe_air$hat_bt50
hat_VT_air <- AHe_air$hat_Vt
qhat_air <- AHe_air$qhat

qhat_list_air <- lapply(1:ncol(qhat_air), function(i) qhat_air[,i
   ])
names(qhat_list_air) <- paste0("qhat", 1:9, "_air")
list2env(qhat_list_air, envir = .GlobalEnv)
```

```
hat_bt0_air <- hat_bt50_air[seq(1, dim)]
hat_bt1_air <- hat_bt50_air[seq((dim + 1), (2 * dim))]
hat_bt2_air <- hat_bt50_air[seq((2 * dim + 1), (3 * dim))]
hat_bt3_air <- hat_bt50_air[seq((3 * dim + 1), (4 * dim))]


order_indices <- order(times, hat_VT_air, qhat_air[,1],
                       hat_bt0_air, hat_bt1_air, hat_bt2_air, hat
                          _bt3_air)

times_ordered <- times[order_indices]
hat_VT_ordered <- hat_VT_air[order_indices]
qhat_ordered_matrix <- qhat_air[order_indices, ]
hat_bt0_ordered <- hat_bt0_air[order_indices]
hat_bt1_ordered <- hat_bt1_air[order_indices]
hat_bt2_ordered <- hat_bt2_air[order_indices]
hat_bt3_ordered <- hat_bt3_air[order_indices]


# Plot coefficient estimators
par(mfrow=c(2,1))

plot(hat_bt0_ordered ~ times_ordered, lwd = 2, type = "l",
     xlab = "Day␣in␣Study", ylab = "Baseline␣Ozone",
     main = "(A)␣Varying␣Intercept␣(Baseline␣Ozone)␣over␣Study␣
        Period");

plot(hat_bt1_ordered ~ times_ordered, lwd = 2, type = "l",
     xlab = "Day␣in␣Study", ylab = "Coefficient␣of␣Solar.R",
     main = "(B)␣Varying␣Effect␣of␣Solar␣Radiation␣over␣Study␣
        Period");

plot(hat_bt2_ordered ~ times_ordered, lwd = 2, type = "l",
     xlab = "Day␣in␣Study", ylab = "Coefficient␣of␣Wind",
     main = "(C)␣Varying␣Effect␣of␣Wind␣Speed␣over␣Study␣Period")
        ;

plot(hat_bt3_ordered ~ times_ordered, lwd = 2, type = "l",
     xlab = "Day␣in␣Study", ylab = "Coefficient␣of␣Temperature",
     main = "(D)␣Varying␣Effect␣of␣Temperature␣over␣Study␣Period"
        );

### Plot variability V(t)
par(mfrow=c(1,1))
ylim_vt <- range(hat_VT_ordered, na.rm = TRUE)
plot(hat_VT_ordered ~ times_ordered, ylim = c(min(hat_VT_air),
   max(hat_VT_air)),
     xlab = "Day␣in␣Study", ylab = "", type = "l", lwd = 2,
     main = "(E)␣Estimated␣Variability␣V(t)␣of␣Ozone␣over␣Study␣
        Period");
mtext(expression(hat(V)(t)), side = 2, cex = 1, line = 3)
```

```
### Plot conditional quantiles estimators
library(ggplot2)
library(tidyr)
library(dplyr)

plot_data_gg <- as.data.frame(qhat_ordered_matrix)
colnames(plot_data_gg) <- paste0("Tau_", taus)
plot_data_gg$DayInStudy <- times_ordered


plot_data_gg_long <- plot_data_gg %>%
  pivot_longer(cols = starts_with("Tau_"),
               names_to = "Quantile_Level",
               names_prefix = "Tau_",
               values_to = "Ozone_Concentration") %>%
  mutate(Quantile_Level = as.factor(as.numeric(Quantile_Level)))


gg_colors <- c("0.1" = "magenta", "0.2" = "gray50", "0.3" = "blue
   ",
               "0.4" = "brown",   "0.5" = "black",  "0.6" = "
                  orange",
               "0.7" = "darkcyan","0.8" = "green",  "0.9" = "red"
                  )
gg_lty <-    c("0.1" = "dashed", "0.2" = "dotted", "0.3" = "
   longdash",
               "0.4" = "dotdash","0.5" = "solid",  "0.6" = "
                  dotdash",
               "0.7" = "longdash","0.8" = "dotted", "0.9" = "
                  dashed")
gg_lwd <-    c("0.1" = 0.8, "0.2" = 0.8, "0.3" = 0.8,
               "0.4" = 0.8, "0.5" = 1.2, "0.6" = 0.8, #
               "0.7" = 0.8, "0.8" = 0.8, "0.9" = 1.2)


p_legend_right <- ggplot(plot_data_gg_long, aes(x = DayInStudy, y
    = Ozone_Concentration,
                                                group = Quantile_
                                                   Level,
                                                color = Quantile_
                                                   Level,
                                                linetype =
                                                   Quantile_Level
                                                   ,
                                                linewidth =
                                                   Quantile_Level
                                                   )) +
  geom_line() +
  scale_color_manual(values = gg_colors, name = "Tau") +
```

```
  scale_linetype_manual(values = gg_lty, name = "Tau") +
  scale_linewidth_manual(values = gg_lwd, name = "Tau") +
  labs(title = "(G) Conditional Quantile Estimators of Ozone (
     extreme day)",
       x = "Day in Study",
       y = "Ozone Concentration (ppb)") +
  theme_minimal() +
  theme(
    legend.position = "right",
    legend.key.size = unit(0.5, "cm"),
    legend.text = element_text(size = 8),
    legend.title = element_text(size = 9, face = "bold"),
    legend.background = element_rect(fill="NA", colour = "NA"),
    legend.box.margin = margin(t = 0, r = 5, b = 0, l = 0, unit =
        "pt"),
    guides(linetype = "none",
           linewidth = "none",
           color = guide_legend(title = "Tau"))
  )
print(p_legend_right)
```

4.2 Dataset

Daily readings of the following air quality values for May 1, 1973 (a Tuesday) to September 30, 1973.

- Ozone: Mean ozone in parts per billion from 1300 to 1500 hours at Roosevelt Island

- Solar.R: Solar radiation in Langleys in the frequency band 4000–7700 Angstroms from 0800 to 1200 hours at Central Park

- Wind: Average wind speed in miles per hour at 0700 and 1000 hours at LaGuardia Airport

- Temp: Maximum daily temperature in degrees Fahrenheit at La Guardia Airport.