Is air pollution correlated to inter-state migration in the US?

EDA Final Project

Leonardo Rueda[1,2] Theo Napitupulu[1,3]

[1] Sanford School of Public Policy [2] jr466@duke.edu [3] yosia.napitupulu@duke.edu

Abstract

This paper does a descriptive analysis of the relationship between air quality and inter-state migration outflows in the United States for the period between 2013 and 2020. Using the Air Quality Databases provided by EPA and the Taxpayer's registries with information about their current and one year before residence location provided by the IRS, we found a positive relationship between the average of days with unhealthy air quality per State and the number of people moving out each state, and between the average of the maximum values registered by the Air Quality Index (AQI) and the number of out-migrants, but we do not find a clear relationship with the number of days with PM2.5 detected in the air. After applying a Pearson's correlation test, we find that the relation in the first two cases is indeed positive and statistically significant, while in the latter, the correlation is not statistically significant. Finally, with a regression analysis, controlling for year and State effects, we found that the only variable with a statistically significant influence in migration is the average number of days with air quality considered unhealthy. Nonetheless, we find a positive relation between air quality and migration, our analysis does not allow us to identify the direction of the effect.

Keywords: Air Pollution, Migration Outflows, Pearson's Correlation Test, Regression Analysis.

Contents

List of Tables

List of Figures

Rationale and Research Questions

This project looks at the effects of air pollution on inter-state migration in the United States using the Air Quality Index datasets from EPA and the Population Migration data from the IRS for the period 2013-2020.

Evidence from middle-income countries shows that air pollution has negative impacts on several health and economic outcomes, such as mortality rates, health expenditures, mental health, hours worked, labor productivity and income. Additionally, other studies have shown how air pollution determines migration decisions. For instance, Chen, S., Oliva, P., & Zhang, P. (2022) found that a 10 percent increase in air pollution, holding everything else constant, reduces population through net outmigration by about 2.8 percent in certain counties in China (see Chen, S., Oliva, P., & Zhang, P. (2022). The effect of air pollution on migration: Evidence from China. Journal of Development Economics, 156, 102833. https://doi.org/10.1016/j.jdeveco.2022.102833).

Our hypothesis is that there is a positive relationship between high air pollution and migration outflows, that is, the highest the pollution registered by the Air Quality Index (AQI) in a state in a year, the higher the number of people leaving that state the same year.

Dataset Information

1. Air Quality Datasets

The Air Quality Index (AQI) dataset provides annual information per county about the maximum values reached by the AQI, the number of days in which this index reached values considered unhealthy, and the number of days with PM2.5 particles recorded. The EPA local air quality stations capture the information. These datasets are available at https://aqs.epa.gov/aqsweb/airdata/download_files.html. We create a dataset with the information from 2013 to 2020 and aggregate the information at the state level. Likewise, we calculate the state averages for each variable.

2. Inter-state migration datasets

The State-to-State outflows dataset provides annual information at the State level about the number of people whose reported home address changed in their individual income tax returns from one year to the other. These datasets are available at https://www.irs.gov/statistics/soi-tax-stats-migration-data. The Inter-state migration datasets do not include the variable "Year", so we create it from 2013 to 2020.

According to the dictionary for this dataset, the variable "y2_statefips" and the code "96" refer to the total outflows of migrants for each state each year. Therefore, we filtered the dataset by that value. Additionally,

we change this variable's name to "FIPS_Code", so later we can merge this dataset with the AQI dataset.

3. Scraping the FIPS codes

The AQI Dataset has the names of each State in the US, but it does not have the code, which is the variable we need to merge this data with the migration one. We scrape the FIPS codes from the webpage (https://www.bls.gov/respondents/mwr/electronic-data-interchange/appendix-d-usps-state-abbreviations-and-fips-codes.htm) and we create a data frame. Then, we merge this dataset with the AQI dataset by the variable "State".

4. Merging the AQI and Migration datasets

We merge and arrange in descending order the AQI and the migration datasets by the variables "FIPS_Code" and "Year". The resulting dataset has information from 2013 to 2020. According to the dictionary for the migration dataset, the variable "n2" refers to the number of individuals who migrated to other states. To facilitate the interpretation, we changed the name of the variable to "Migrants.outflows". We create a subset of the previous dataset with the variables of interest: FIPS_Code, Year, State, Avg.Unhealthy.DAys, Avg.Days.AQI, Avg.Days.PM2.5, and Migrants.outflows. This is the dataset that we will use in our analysis.

The final dataset has seven variables: FIPS Code, Year, State, Avg.Unhealthy.Days, Avg.Max.AQI, Avg.Days.PM2.5, and number of migrants. In table 1 we can see the characteristics of the dataset for years 2013 and 2020. Each year has information for the fifty states of the USA (we do not include in the analysis the District of Columbia, Puerto Rico and the Virgin Islands). For 2013, on average each state showed 0.56 days with air quality classified as "unhealthy", the maximum value reached by the AQI was 114.89, the number of days with PM2.5 registered were 116.57, and 135,657 people migrated to a different state. For 2020, on average each state showed 0.90 days with air quality classified as "unhealthy", the maximum value reach by the AQI was 122.81, the number of days with PM2.5 registered were 121.52, and about 138,954 people migrated to a different state.

Table 1: Summary statistics for the final dataset 2013-2020

Variable	Type	Obs	Min	Median	Media	Max
Year=2013						
FIPS Code	int	50				
State	chr	50				
Avg.Unhealthy.Days	num	50	0.00000	0.03348	0.56471	6.13333
Avg.Max.AQI	num	50	73.00	107.22	114.89	240.85
Avg.Days.PM2.5	num	50	17.43	110.33	116.57	309.00
Migrants.outflows	num	50	16540	95343	135657	532619
Year=2020						
FIPS Code	int	50				
State	chr	50				
Avg.Unhealthy.Days	num	50	0.00000	0.09762	0.90793	14.20755
Avg.Max.AQI	num	50	58.75	98.25	122.81	430.35
Avg.Days.PM2.5	num	50	16.11	105.87	121.52	304.39
Migrants.outflows	num	50	16378	92877	138954	684935

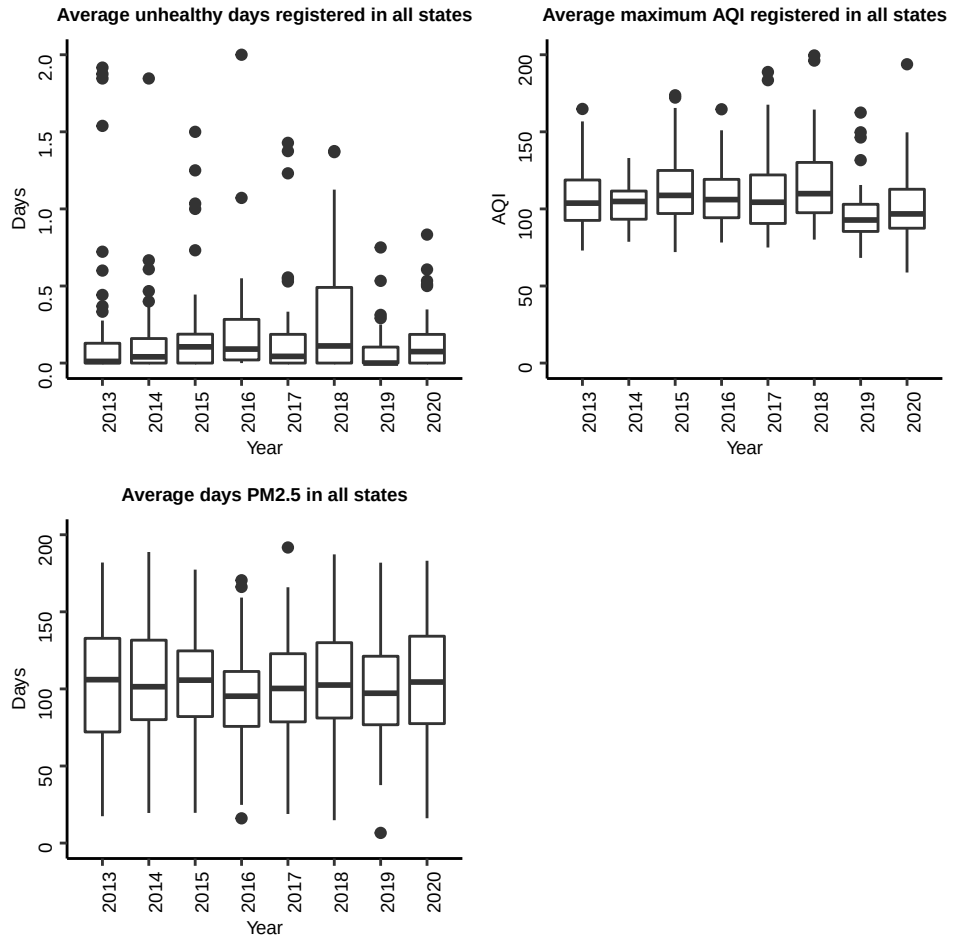
Exploratory Analysis

Air Quality Indicators between 2013 and 2020

In figure 1 we can see some descriptive statistics for the air quality indicators between 2013 and 2020. The graph at the top left, the average unhealthy days registered in all states and for all years, shows no days with unhealthy air quality. However, between 2013 and 2016, some states registered on average of two days, and between 2018 and 2020, these registries dropped to one day. In the graph at the top right, the average maximum AQI registered for all states reached numbers close to one hundred, which means that the air quality is acceptable. At the same time, in some years this indicator reached numbers above between 150 and 200, meaning that the air quality was poor, and people could experience negative health effects. Finally, the graph at the bottom left shows the average number of days when PM2.5 were detected, which reached one hundred during the period analyzed. This result is worrisome because prolonged exposure to these particles harms human health. This indicator shows a high dispersion, with some states showing less than

50 days, while others show more than 150 days. It is important to mention that the analysis of averages by state per year allows having an initial picture of the air quality in each year but does not allow to identify the air quality by state. In the following analysis, some maps allow to see this.

Figure 1: Air Quality Indicators 2013-2020

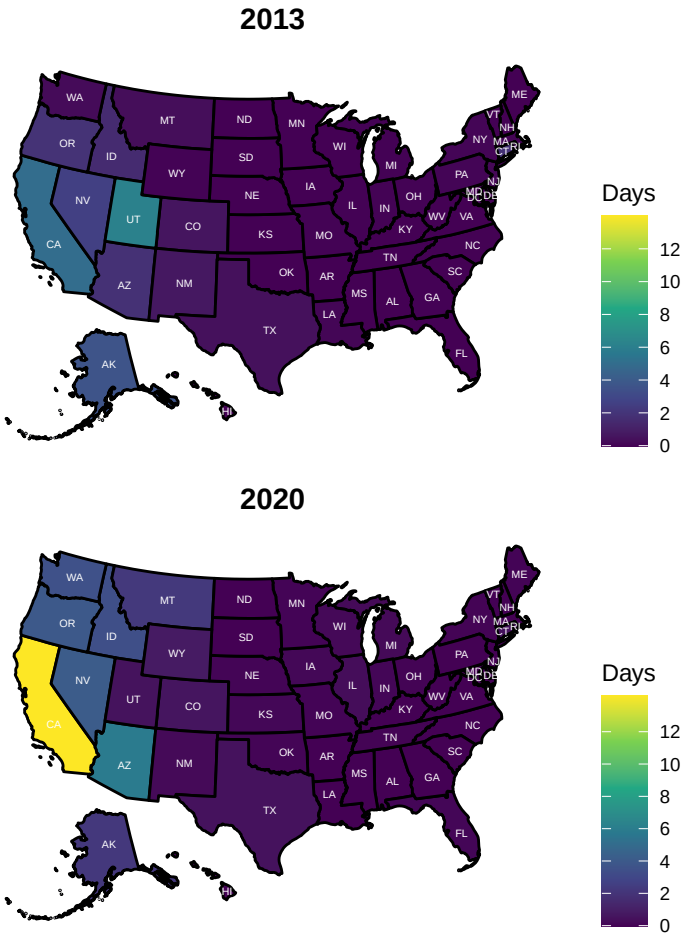


Average Unhealthy Days by State between 2013 and 2020

Figure 2 shows how the indicator "average unhealthy days" developed between 2013 and 2020 by State. During 2013, most states had an average unhealthy days indicator close to zero, except states close to the west coast such as Utah (6.13), California (5.0), Alaska (3.62), Nevada (2.60), Oregon (1.91), Connecticut (1.87), Arizona (1.84) and Idaho (1.53). In 2020, the picture is similar, but the average number of unhealthy days worsens in the states on the west coast. In particular, California and Arizona more than double their average unhealthy days indicator (14.2 and 5.84, respectively), while Nevada (4.11), Oregon (3.95) and Idaho (3.40) also worsen their registries. Further, Washington and Montana show significantly higher numbers

compared to 2013 (3.54 and 2.26, respectively). On the other hand, Connecticut and Utah significantly reduced their average unhealthy days indicator (0.50 and 0.53 respectively).

Figure 2: Average Unhealthy Days by State 2013 and 2020

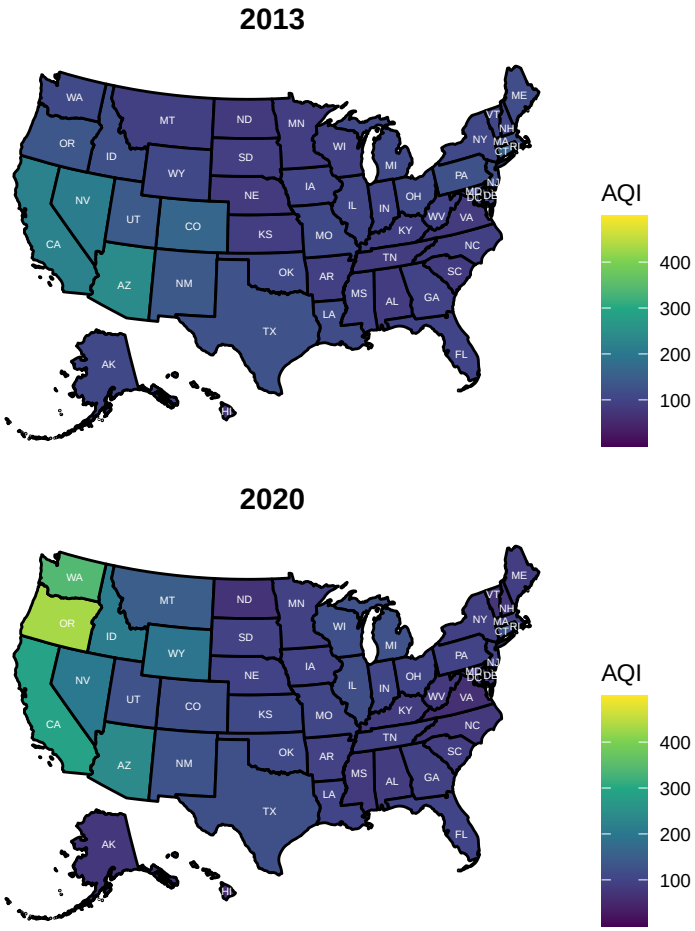


Average Maximum AQI Values by State between 2013 and 2020

Figure 3 shows the evolution of the indicator "average maximum AQI value" between 2013 and 2020 by State. In 2013, only 34 percent of states show an average maximum AQI value below one hundred, which, according to EPA, ranges between good and moderate air quality and most people can enjoy outdoor activities (except unusually sensitive people). The remaining 66 percent states show numbers above 100, which means that the air quality is unhealthy for sensitive groups (people with heart or lung decrease, for instance) and among them, six states show numbers above 150 which represents a health threat for everyone. Those states were Arizona (240), California (219), Nevada (209), Colorado (164), Connecticut (156) and Rhode Island (154). In 2020, 54 percent of states presented Average Maximum AQI values below one hundred, which means that a

little over half of states improved their registries. On the contrary, seven states reached numbers significantly higher and in a couple of cases hazardous for everyone. Those states were Oregon (430), Washington (340), California (287), Arizona (238), Idaho (209), Nevada (202) and Wyoming (193).

Figure 3: Average Max AQI value by State 2013 and 2020

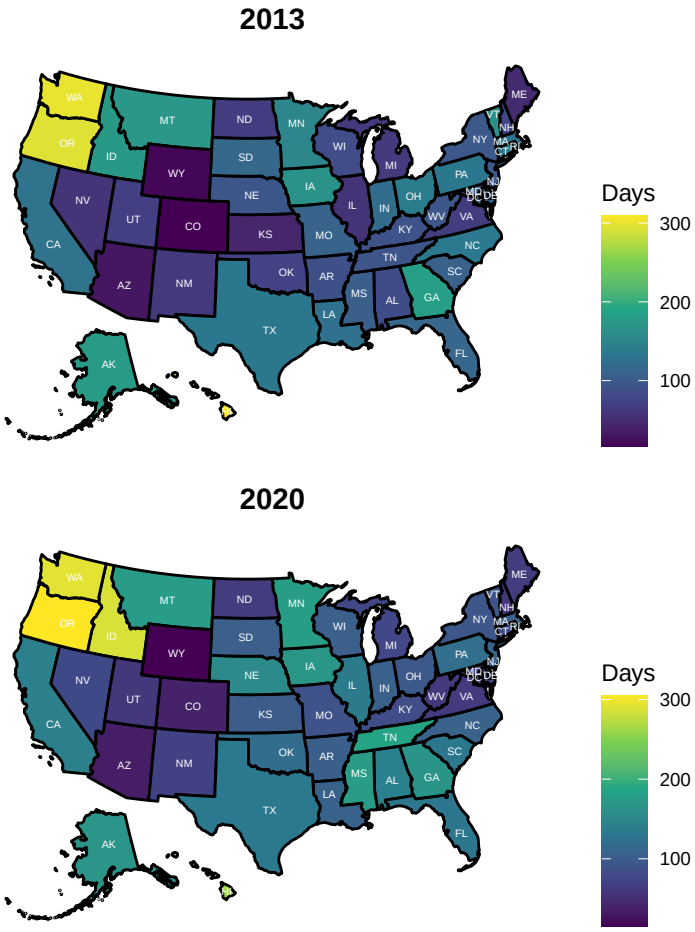


Average Days with PM2.5 by State between 2013 and 2020

Figure 4 shows the indicator "average days with PM2.5 registered" between 2013 and 2020 by State. For 2013, 36 percent of the states spent on average at least one quarter during that year with this type of particle registered in the air, while 60 percent spent at least two quarters. Three states spent a little over three-quarters with this type of particles registered in the air, specifically Hawaii (309), Washington (299) and Oregon (294). In 2020, 32 percent of states spent less than two quarters with this type of particles registered in the air, while the same 60 percent spent at least two quarters. However, five states spent over two and sometimes three quarters with this type of particles detected in the air. These states were Oregon

(304), Washington (291), Idaho (209), Hawaii (262) and Tennessee (183).

Figure 4: Average PM2.5 by State 2013 and 2020



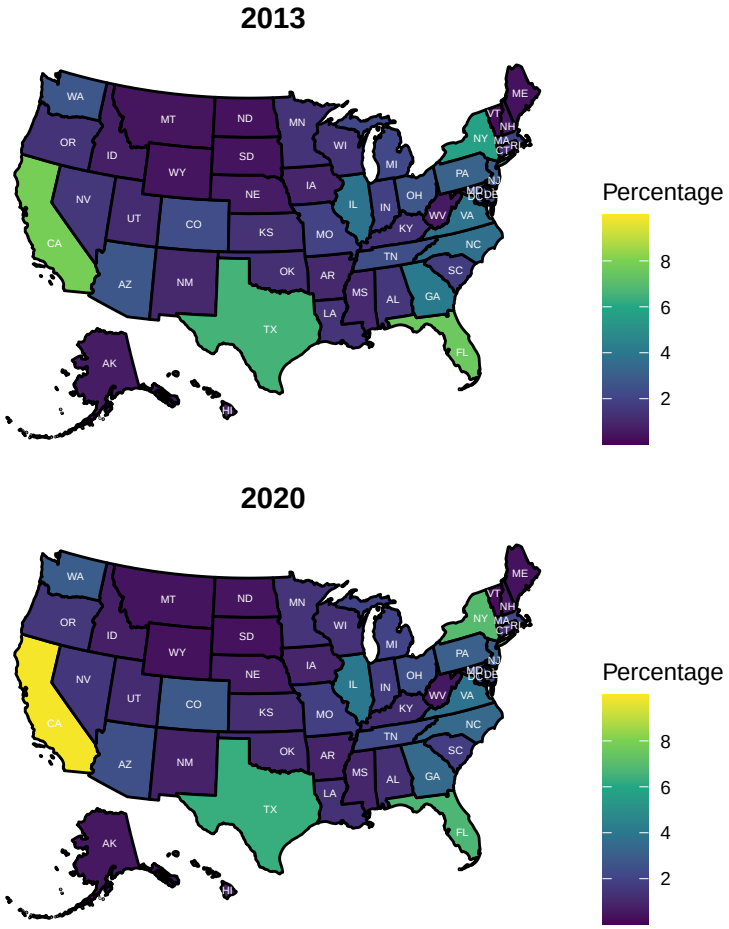
It is important to mention that these results should be interpreted with caution because the state averages do not allow us to understand what the distribution per county inside each state is. It could be the case that air quality registries in one county over influence the average for the entire state, or at the other extreme, maybe the air quality is equally distributed in each state.

Migration outflows by State between 2013 and 2020

Figure 5 shows the proportion of migrants leaving each State in 2013 and 2020. During the first year, eleven States explained 52 percent of the migration outflows. Those States were California (7.85%), Florida (7.54%), Texas (6.52%), New York (5.69%), Georgia (4.07%), Illinois (3.82%), Virginia (3.80%), North Carolina (3.67%), Pennsylvania (3.16%), New Jersey (2.91%) and Arizona (2.73%). During 2020, most of the same States explain the migration outflows with some minor changes. In specific, California explains

9.85% of all migration outflows, followed by New York (6.90%), Florida (6.70%), Texas (6.29%), Illinois (3.94%), Virginia (3.76%), Georgia (3.45%), North Carolina (3.41%), Pennsylvania (3.07%), and New Jersey (3.01%).

Figure 5: Percentage of migrants' outflows by State in 2013 and 2020



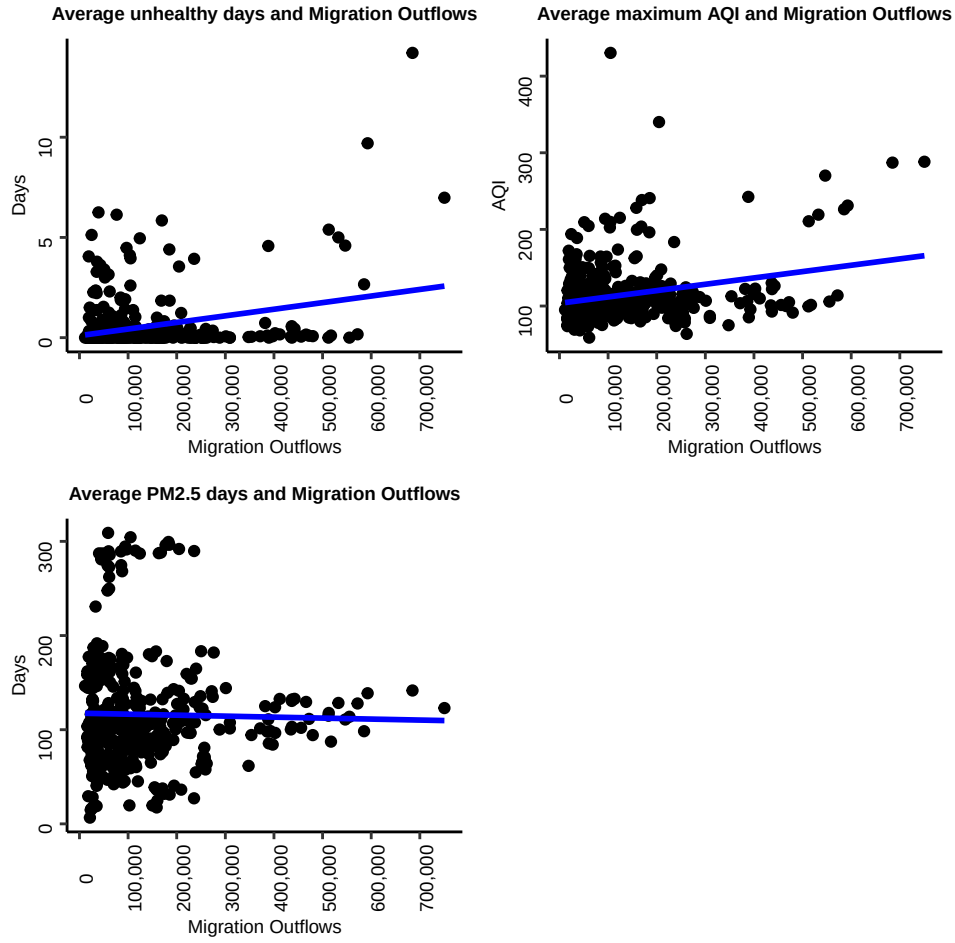
Analysis

Air Quality and migration outflows

In figure 6 we can see the relationship between each of the air quality indicators and the migration outflows for the seven years studied. The top left graph shows a positive relationship between the average unhealthy days and the number of people out-migrating from a particular state: the larger the number of days considered unhealthy, the larger the number of people moving out to another state. The figure at the top right also shows a positive relationship between the average maximum AQI value and migration outflows: the larger the AQI maximum values reached in each state, the higher the number of more people leaving that state.

Finally, at the bottom left, the figure does not show a clear relationship between the average PM2.5 days and migration outflows (at first sight it seems to be negative, but it is difficult to tell).

Figure 6: Relationship between air quality indicators and migrantion outflows between 2013 and 2020



Correlations

To confirm the type of statistical relationship between the air quality indicators and the migration outflows, we run some Pearson's correlation tests for these variables. Table 2 confirms that the correlation between average unhealthy days and migration outflows per State is positive ($\rho = 0.29$) and statistically significant (p-value = 2.609e-09). Likewise, the average maximum AQI value and the migration outflows per State show a positive ($\rho = 0.25$) and significant (p-value = 1.558e-07) correlation. However, it is not conclusive the type of relationship between the average of days with PM2.5 registered and the migration outflows, because the interval at 95% of confidence falls between negative and positive values and the relationship is not statistically significant.

Table 2: Correlation between air quality indicators and migration outflows 2013-2020

Variables	Migrants Outflows	t	p-value	Min	Max
Avg. unhealthy days	0.2921202	6.0936	2.609e-09	0.1997911	0.3793075
Avg. maximum AQI	0.2586106	5.341	1.558e-07	0.1647361	0.3478420
Avg. days PM2.5	-0.02109288	-0.4209	0.6741	-0.118899	0.07711834

Regression Analysis

To understand better the true effect of air quality in the inter-state migration outflows, we run a linear regression with the following specification:

$$Mig.out = \alpha + \beta_1 * Avg.unhealth.days + \beta_2 * Avg.max.AQI + \beta_3 * Avg.days.PM2.5 + \beta_4 * Year + \beta_5 * State + \epsilon$$

We assume that migration is explained by the average unhealthy days, the average maximum AQI value, the average days with PM2.5, and we include some Year and State Dummies to control for any effect per year or State.

Table 3 shows the results. In general, the model predicts 94% percent of the variability of migration outflows, which is a high value given that there are many other economic and social variables that motivates migration that we are disregarding (though some of these effects may be partially capture by the Year and State control variables). The high R-Squared could also result from a high correlation between the explanatory variables (which is possible because all express the same phenomenon) and means that the model is highly biased.

The results also indicate that the only variable highly statistically significant to explain migration outflows is the Average Unhealthy Days per State, and the coefficient suggests that when the average unhealthy days increase by one unit, about 7,535 people move out to a different State. On the other hand, the variables Average Maximum AQI value and Average Days with PM2.5 are not significant.

Table 3: Regression Analysis - migration outflows and air quality indicators

Migration Outflows	*Coefficient*	*S.E.*	*p-value*
Intercept	-3688365	1306551	0.01
Avg.Unhealthy.Days	7535.74	2104.64	0.00
Avg.Max.AQI	-20.70	70.14	0.77
Avg.Days.PM2.5	-17.46	88.63	0.84
R-squared	0.95		
Adj.R-squared	0.94		
F	124.32		0.00
N	400		

It is important to mention that this analysis is not conclusive, because it does not allow us to infer anything about causality. We assume that air quality causes migration, but the other way around could also be true. To address endogeneity problems, a more sophisticated analysis is needed.

Summary and Conclusions

This project does a first approximation to the relationship between inter-state migration and air quality in the United States. Apparently, the causal relationship between air quality and migration has not been explored enough yet because of the empirical challenges it imposes to researchers (see Chen, S., Oliva, P., & Zhang, P. (2022), p.2). Using the air quality information provided by EPA and data about taxpayer's current state of residence, we analyze if bad air quality indicators are related to high migration outflows during the period 2013 and 2020. Through a descriptive analysis, we found that there is a high dispersion of air quality throughout the United States, with some states performing constantly bad (most of them along the west coast) while others seem to slightly improve or remain the same. A visual analysis allows us to discover a positive relationship between the number of days with air quality considered as unhealthy and the migration outflows, and the average maximum AQI value and the migration outflows (not in the case of the average PM2.5 days and the migration outflows). Then, we applied the Pearson's correlation test and found that indeed the relationship is positive and significant for the cases aforementioned. Finally, through a linear regression, after removing year and State effects, we discover that the only statistically significant variable

affecting the migration outflows is the average unhealthy days. However, these results must be interpreted with caution. First, taking the averages by State for the air quality indicators is misleading because it disregards the distribution of the indicator inside each State. Second, the regression analysis performed does not eliminate the double causality problem that we face. Third, a better approach would be to take the net migration outflows, instead of the number of people just leaving each State.