Spatial and Temporal Patterns in Health Code Violations in LA County

Albert Kuo

3/6/2019

1 Abstract

We conduct an analysis of the frequency and pattern over time and location of health code violations from 2015 to 2017 in the Los Angeles county. We find that there are significant differences in violation rates for a few health codes between cities in the LA county. In addition, we identify several distinct temporal patterns in health code violations. Some are due to apparent changes in health code definitions, while other health codes exhibit seasonal variation or sustained increases in the number of violations.

2 Introduction

Restaurant cleanliness is an important facet of public health, particularly for consumers who frequently eat out. To administer standards of hygiene and cleanliness, restaurants are routinely investigated by health inspectors. Systematic patterns in health code violations are of interest to policymakers as well as the general public. For instance, what are the most common types of health code violations? Do certain areas have higher violation rates? Are health code violations increasing over time? With the recent release of health code violations data in the LA county on Kaggle, we can begin to answer these questions in order to inform future policies.

3 Methods

3.1 Standardization of Rates

In the spatial and temporal exploratory analyses, we use standardized rates to more easily compare different health codes visually. In the spatial analysis, the standardized rate $p_{v,s}$ is defined as $p_{v,s} = \left(\frac{n_{v,s}}{m_s} - \mu_v\right)/\sigma_v$ where $n_{v,s}$ is the number of violations for health code v in facility city s, m_s is the total number of inspections in facility city s, and μ_v and σ_v is the mean and standard deviation of the rate of violations for a health code v.

Similarly, in the temporal analysis, the standardized rate $p_{v,t}$ is defined as $p_{v,t} = \left(\frac{n_{v,t}}{m_t} - \mu_v\right)/\sigma_v$ where $n_{v,t}$ is the number of violations for health code v on day t, m_t is the total number of inspections on day t, and μ_v and σ_v are defined as above.

3.2 Loess Smoothing

To explore long-term temporal trends in the number of health code violations, we use local polynomial regression fitting (loess). In loess, the fit to the data at a point x is estimated using points in a local neighborhood around x, weighted by their distance from x (1). This allows the fitted curve to reflect smooth, non-linear temporal patterns in the data. By fitting a unique loess curve for each type of health code violation, those that follow an unusual trend over time can be distinguished visually and further analyzed. Due to large differences in the number of violations per health code, loess curves are fitted to standardized rates so that different health codes can be more easily compared.

3.3 Log-linear Model

A log-linear regression model is used to assess the statistical significance of spatial trends in violation rates by city: $\log(n) = x_1 + \log(x_2)$. Each observation corresponds to a city. The outcome n_i is the number of violations in city i and the primary independent variable is an indicator variable for a specified city (or cities) of interest (x_1), adjusted for the total number of inspections in city i (x_2). Each type of health code violation of interest is regressed separately to analyze health code specific spatial patterns. It is assumed that:

1. The observations are independent between cities.
2. The residuals are distributed normally (checked with model diagnostics)
3. The residuals exhibit constant variance (checked with model diagnostics).

3.4 Logistic Regression

Logistic regression is used to assess the statistical significance of temporal trends. Each facility in the data is assumed to be independent, that is, the number of health code violations at one facility do not affect the number of violations at a different facility. Thus, the total number of health code violations across all facilities on a given day i can be modeled by a binomial distribution $Binomial(n_i, \pi_i)$, where n_i is the total number of facilities and π_i is the probability that a facility gets a health code violation. In logistic regression, the relationship between the logit of π_i and the independent variables x_i is assumed to be linear.

$$\text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = x_i^T\beta \tag{1}$$

$$= \beta_0 + \beta_1 * x_1 + (x_i^*)^T\beta^* + \beta_2 * x_2 + \beta_3 * x_3 \tag{2}$$

In the temporal analysis models, we test whether violations change over time. The primary independent variables are the activity date (x_1) and additional terms to capture nonlinear trends in time (x_i^*), adjusted for the total number of inspections on that date (x_2) and the day of the week (x_3). Apparent nonlinear trends in time (x_i^*) may be modeled using linear spline terms, with knots are determined through exploratory visual analysis, or by using the month and season for seasonal trends. Each type of health code violation of interest is regressed separately to analyze health code specific temporal patterns.

4 Results

4.1 Data Characteristics

From June 30, 2015 to December 28, 2017, a total of 43,654 facilities in the LA county were inspected for health code violations. In the 190,828 health inspections conducted, 906,014 health code violations were recorded. Out of 906,104 health code violations, 8,640 (1%) are missing metadata, including the date and location, and dropped from the analysis.

The ten most common health code violations, out of a total of 112 health codes, are listed in Table 1. They account for 65% of all health code violations. Some violation codes are very rare and their data is sparse (Supplementary Figure 10). 54 health codes are of Type "F" (begin with the letter "F") and these violations make up 99.9% of all violations. Therefore, the analysis is restricted to only Type "F" violations.

Table 1: Top 10 Most Common Violations

Violation code	Violation	N	Percent
F044	Floors, walls and ceilings: properly built, maintained in good repair and clean	102012	11.3%
F033	Nonfood-contact surfaces clean and in good repair	100083	11.0%
F035	Equipment/Utensils - approved; installed; clean; good repair, capacity	80020	8.8%
F040	Plumbing: Plumbing in good repair, proper backflow devices	50870	5.6%
F036	Equipment, utensils and linens: storage and use	49744	5.5%
F037	Adequate ventilation and lighting; designated areas, use	48046	5.3%
F043	Premises; personal/cleaning items; vermin-proofing	42949	4.7%
F007	Proper hot and cold holding temperatures	41110	4.5%
F030	Food properly stored; food storage containers identified	39855	4.4%
F039	Wiping cloths: properly used and stored	35849	4.0%

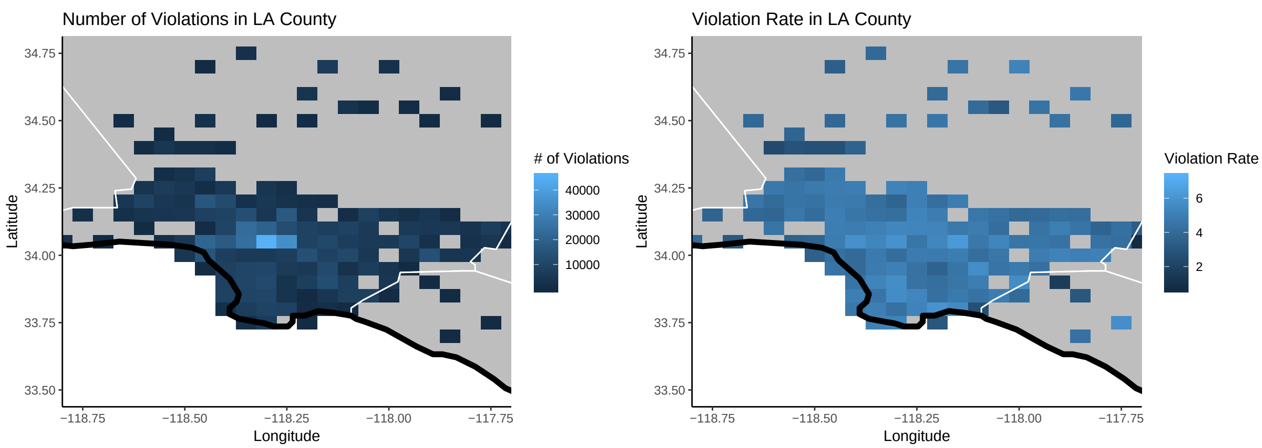

Figure 1: Map of violations (left) and violation rate (right). The spatial pattern in the number of violations is not reflected in the violation rate.

4.2 Exploration of Spatial Patterns in Health Code Violations

In Figure 1, the total number of violations and violation rate (number of violations/number of inspections) for each spatial block is plotted on a map of LA county. While there is a clear spatial pattern in the number of violations, this pattern is not apparent in the violation rate. In other words, after controlling for the number of inspections, the rate of health code violations are comparable across the region overall.

To investigate spatial patterns for specific health codes, we use a heat map of standardized violation rates (Figure 2). Spatial patterns by city can be discerned by variation along the y-axis. If there are no differences in the violation rate between cities for a given health code, then all the tiles in the corresponding column will have the same color. Differences in spatial patterns between health codes can therefore be observed by comparing their respective columns. Moreover, if a city has a relatively low violation rate for all health codes, then its corresponding row will be consistently lighter in color.

Overall, there are few spatial patterns in the most common violations. The clearest signals include Santa Monica, which appears to have a lower violation rate for the two most common health code violations: (1) Conditions of floors, walls, and ceilings (F044) and (2) Clean surfaces (F033). In addition, Avalon appears to have higher violation rates for a number of health codes, particularly the health code for the presence of rodents, insects, and other animals (F023). Santa Monica also appears to have a higher violation rate for F023. These exploratory observations are further investigated in the next two sections.

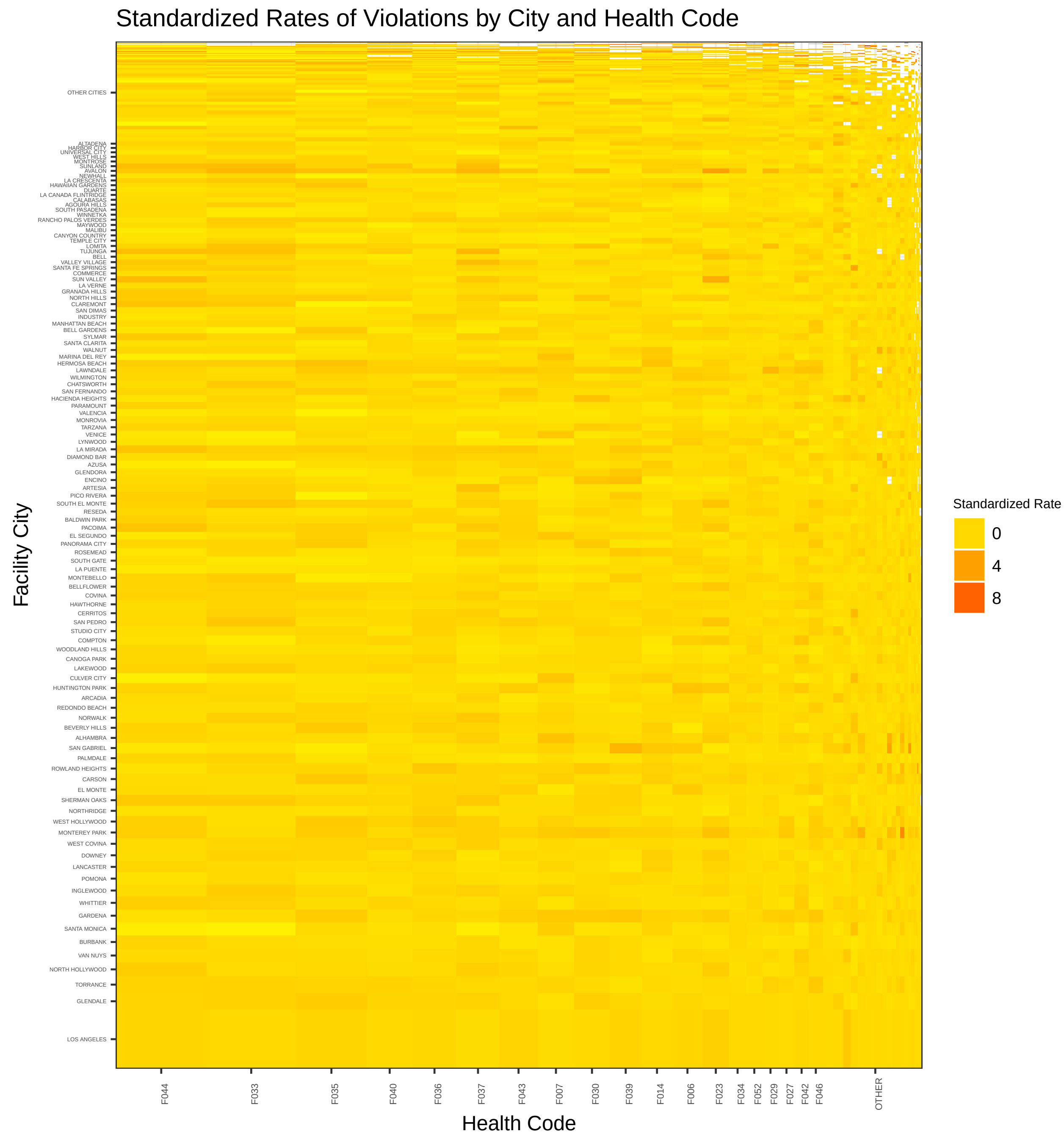

Figure 2: Heat map of standardized violation rates by city and health code. The heights of the tiles are proportional to the number of violations for the respective city and the widths of the tiles are proportional to the number of violations for the respective health code. Both axes are ordered by the number of violations. If there are no recorded violations for a particular health code and city combination, then the corresponding tile appears as a blank white tile.

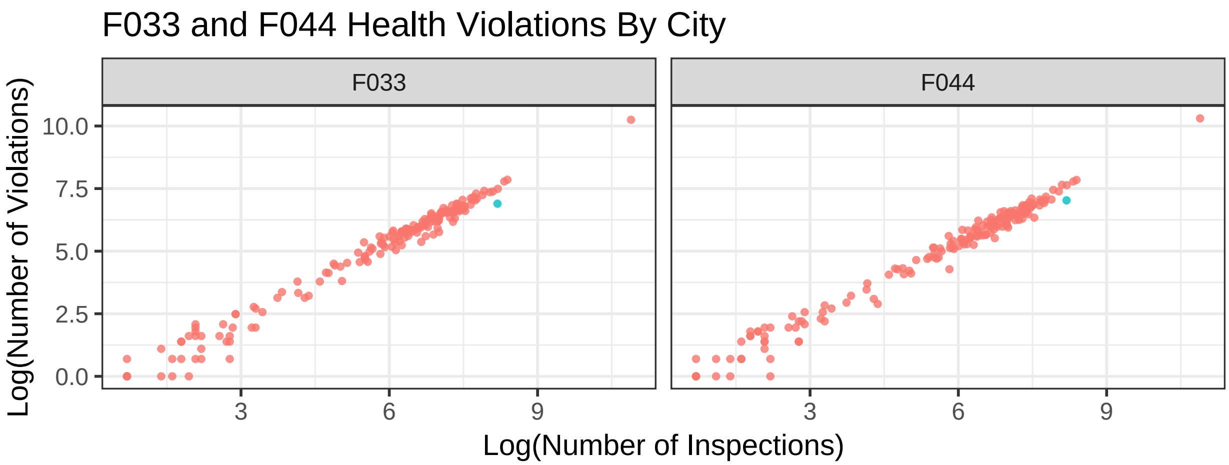

Figure 3: Santa Monica (green points) has lower violation rates for both health codes F033 and F044.

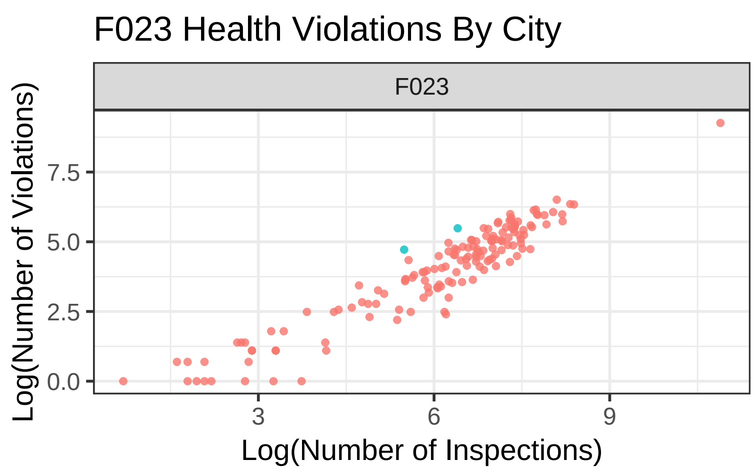

Figure 4: Avalon and Sun Valley (green points) have higher violation rates for health code F023.

4.3 Lower F033 Violation Rates in Santa Monica

Violation rates in Santa Monica are lower than those in other cities for the health codes for conditions of floors, walls, and ceilings (F044) and clean surfaces (F033), as seen in Figure 3. Using a log-linear regression model, this difference is significant for F033 (p-value of 0.039 under a two-sided t-test with Bonferroni correction), but not F044 (p-value of 0.16 under a two-sided t-test with Bonferroni correction). Based on model diagnostics, outlier cities are removed (three cities and one city for F033 and F044 respectively) and in the resulting models, assumptions are not severely violated (Supplementary Figures 11 and 12).

4.4 Higher F023 Violation Rates in Avalon and Sun Valley

Avalon and Sun Valley have higher violation rates for the presence of rodents, insects, and other animals (F023), as seen in Figure 4. Using a log-linear regression model, this difference between the two cities and the rest of the cities is significant (p-value of 0.002 under a two-sided t-test). Model assumptions are not severely violated (Supplementary Figure 13).

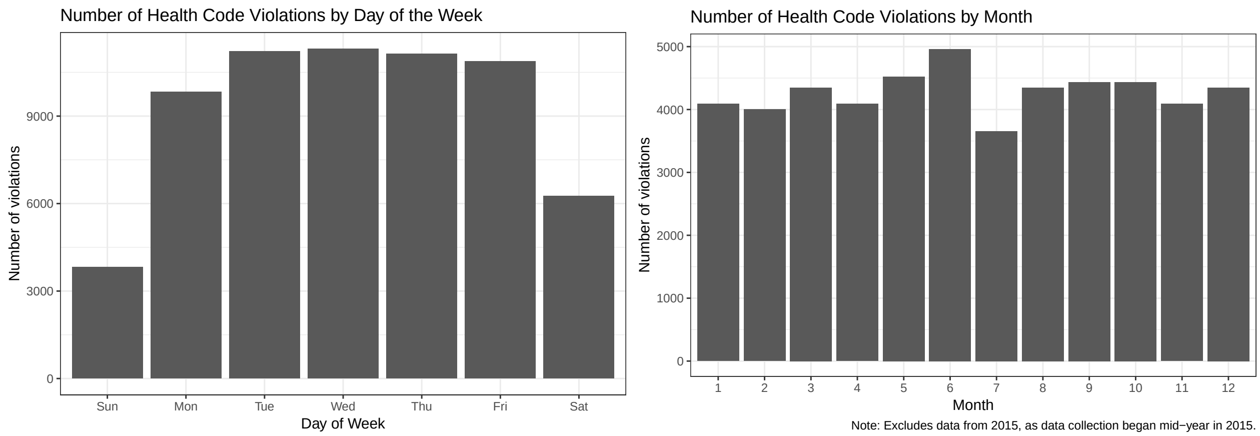

Figure 5: Overall, health code violations are more common during weekdays and do not exhibit seasonal patterns.

4.5 Exploration of Temporal Patterns and Trends in Health Code Violations

As expected, health code violations are more common during the weekday, corresponding to the likely work hours of health code inspectors (Figure 5). Due to this apparent pattern, long-term time trends are adjusted for the day of the week in the logistic regression models. There are no clear seasonal patterns in the number of overall health code violations, though July appears to have a slightly lower count than expected.

To explore long-term trends for specific health codes, loess curves are fitted to standardized counts and plotted for every type of health code violation (Supplementary Figure 14). Seven health codes of interest are subsequently identified:

1. F021 - Water availability
2. F023 - Presence of rodents, insects, and other animals
3. F048 - Permits available
4. F049 - Impoundment of unsanitary equipment or food
5. F053 - Hot water availability
6. F054 - Multiple major critical violations
7. F058 - Raw or undercooked foods

These health codes display one of the three temporal trends shown in Figure 6. We analyze the significance of the trends and investigate potential causes in the next sections.

4.6 Changes In Health Codes Over Time

Some of the most significant temporal shifts occur due to changes in how health code violations are recorded. For example, after 2017, the number of water availability (F021) health code violations drops significantly, most of which can be accounted for by the emergence of hot water availability (F053) health code violations (Figure 7). While a few violations of water availability (F021) are recorded after 2017, they are rare compared to the number of violations of hot water availability, which are recorded at a frequency comparable to that of water availability violations prior to 2017. This suggests that there was a change in the health code used to record this violation.

A similar pattern can be observed in the multiple major critical violations health code (F054). Prior to 2017, there were no recorded violations for this health code (Figure 7). Unlike with the water availability health code, no corresponding similar health code could be found, which suggests that the multiple major

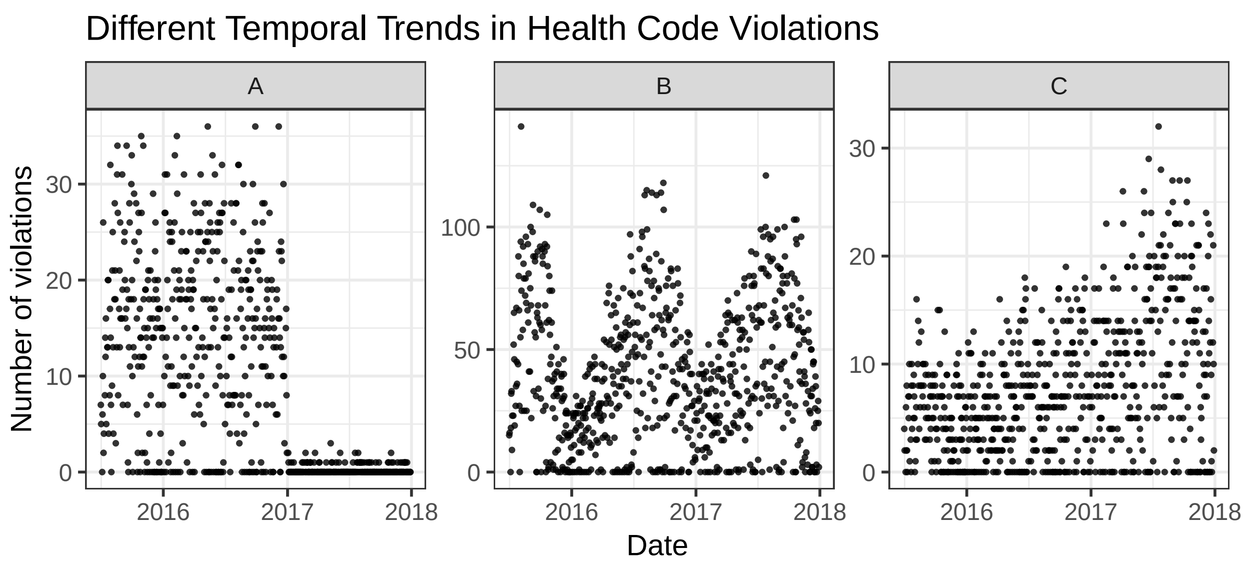

Figure 6: Three types of temporal trends can be observed in health code violations.

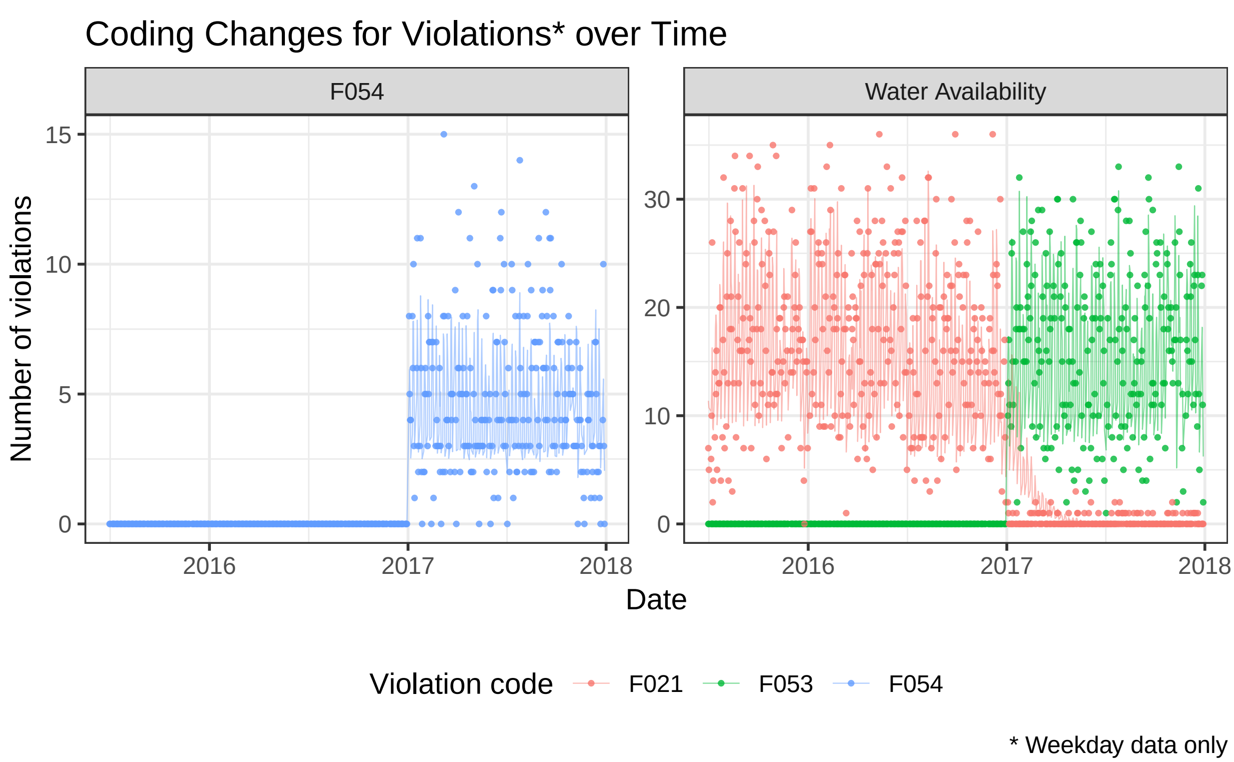

Figure 7: Left: The multiple major critical violations health code (F054) appears to have been introduced in 2017. Right: The number of water availability health code (F021) violations dropped in 2017, when the hot water availability health code (F053) was introduced. Fitted lines are estimated using logistic regression, including a linear spline term with a knot at 2017.

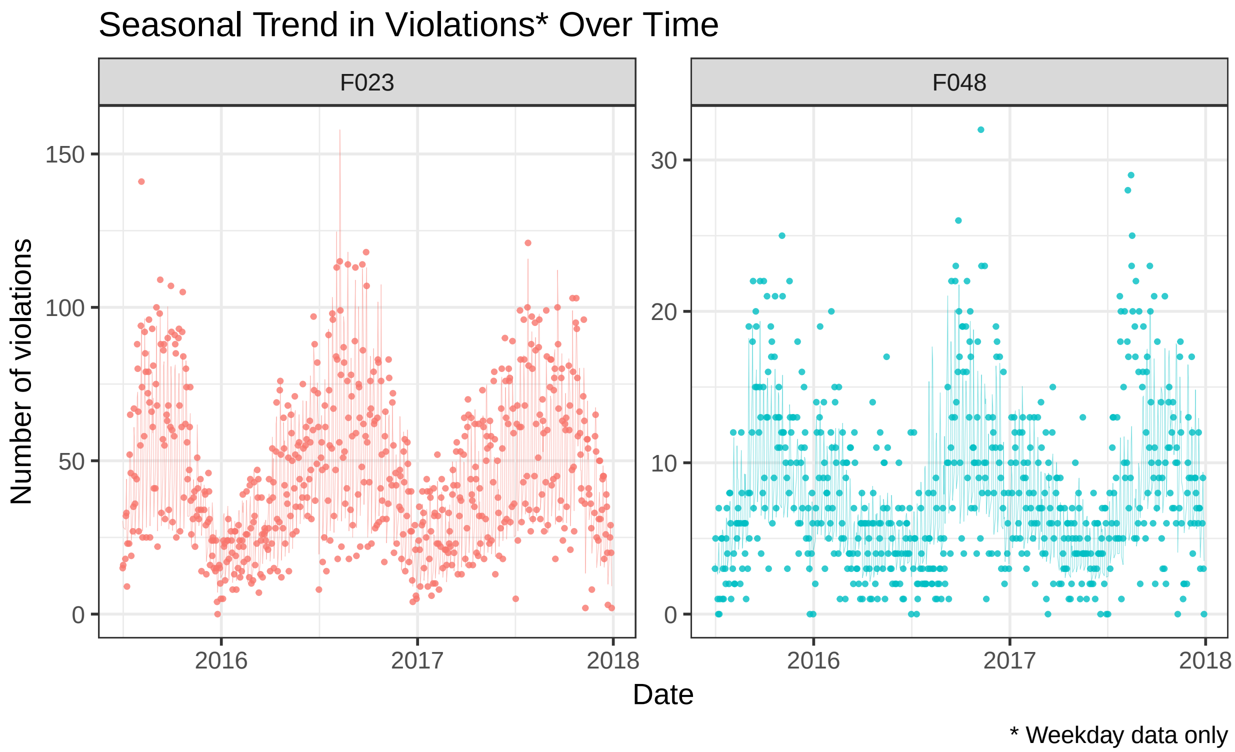

Figure 8: Seasonal patterns in F023 and F048 are shown. Fitted lines are estimated using logistic regression, including month as a categorical independent variable.

critical violations health code was a newly introduced health code in 2017, rather than a replacement for a pre-existing health code.

4.7 Seasonal Trends

Violations for the presence of rodents, insects, and other animals health code (F023) exhibit a seasonal pattern, with higher number of violations over the summer and lower number of violations during the winter (Figure 8). This is in agreement with the fact that rodents and pests are more common during the summer season. A seasonal trend can also be observed in violations for permit availability (F048). This may correspond with renewal schedules in permits or a rise in pop-up establishments without permits during the latter half of the year. For both health codes, the month terms are significant using a drop-in deviance test (p-value < 0.01), providing evidence for a seasonal trend.

4.8 Increases In Violations Over Time

In Figure 9, we observe an increase in the number of violations for the impoundment of unsanitary equipment or food (F049) and to a lesser extent, raw or undercooked foods (F058). The linear time trends are significant for both (p-value < 0.01 under a two-sided t-test). The lack of a similar trend in other types of violations merits further investigation into why these particular health code violations, particularly F049, have increased in number over the years.

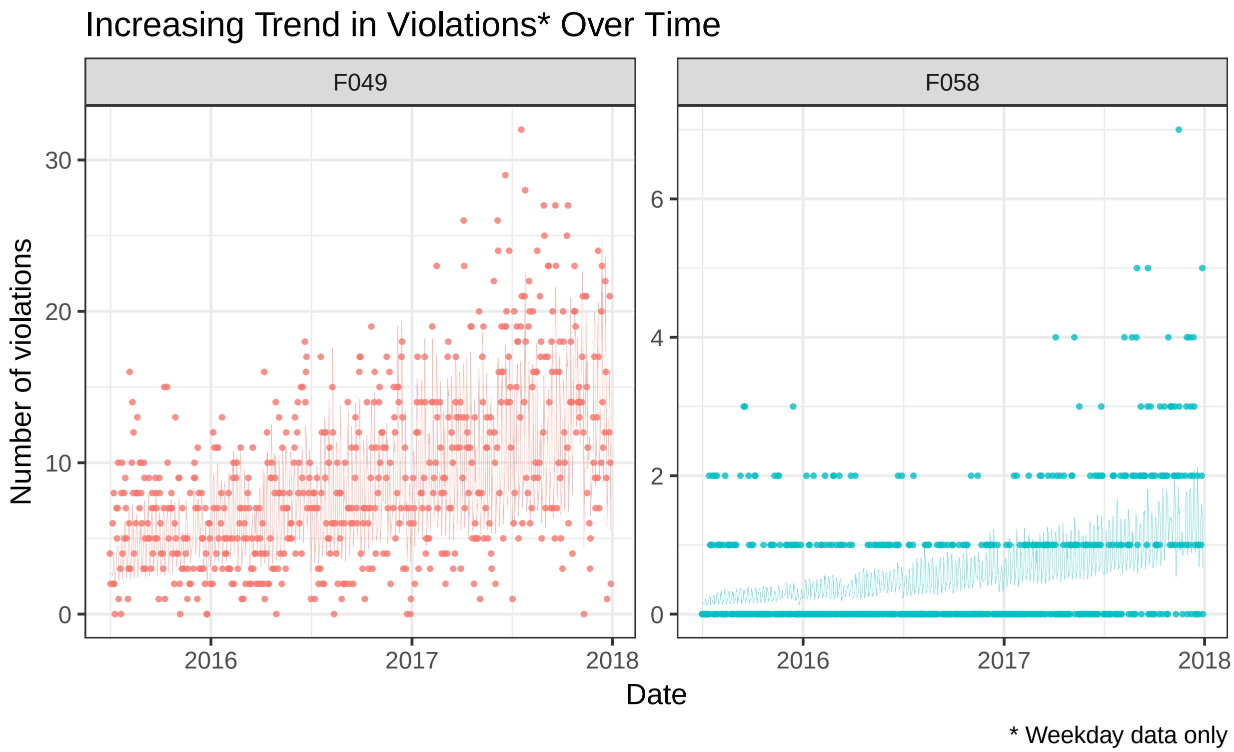

Figure 9: Increasing trends in F049 and F058 are shown. Fitted lines are estimated using logistic regression.

5 Discussion

Over 100 different types of health code violations and 54 Type "F" violations were recorded in the LA county during a 2.5-year period from 2015 to 2017. However, most are violations of only a handful of health codes, with the ten most common health code violations accounting for 65% of all violations (Table 1) and the 54 Type "F" violations accounting for 99.9% of all violations.

While there are no clear spatial trends in the overall violation rate (Figure 1), there is evidence for some significant health code specific differences in violation rate among cities (Figure 2). In particular, Santa Monica has a statistically significant lower violation rate for clean surfaces (F033). Moreover, Avalon and Sun Valley have significantly higher violation rates for the presence of rodents, insects, and other animals (F023). These observed patterns may reflect a true difference in hygiene standards between cities, but may also reflect city-specific differences in the stringency of investigations.

Temporally, clear patterns in the number of health code violations emerge according to the day of the week, likely corresponding to the work schedule of health investigators (Figure 5). In addition, we observe three distinct long-term temporal trends in health code violations (Figure 6). Some of these observed temporal trends can be attributed to changes in how health codes are recorded, such as in the case of the water availability health code (Figure 7). Others, such as the presence of rodents, insects, and other animals (F023) can be explained by seasonal ecological trends (Figure 8). However, some health codes, such as the impoundment of unsanitary equipment or food (F049), appear to be increasing over time, with no current known causes (Figure 9). While it's possible that a long-term increase in these violations reflects a decline in sanitary conditions in facilities over time, the lack of a similar trend in other health codes indicate that the causes are specific to these health codes. For instance, the recording or violation of certain health codes over others may have been incentivized differently over time due to policy changes in fines or other regulations. Further investigation into why the violation rates of these health codes are increasing may be of interest to policymakers.

6 References

1. Cleveland, William S. "Robust locally weighted regression and smoothing scatterplots." Journal of the American statistical association 74.368 (1979): 829-836.

!"##$%&%'()*+,-./"*%0

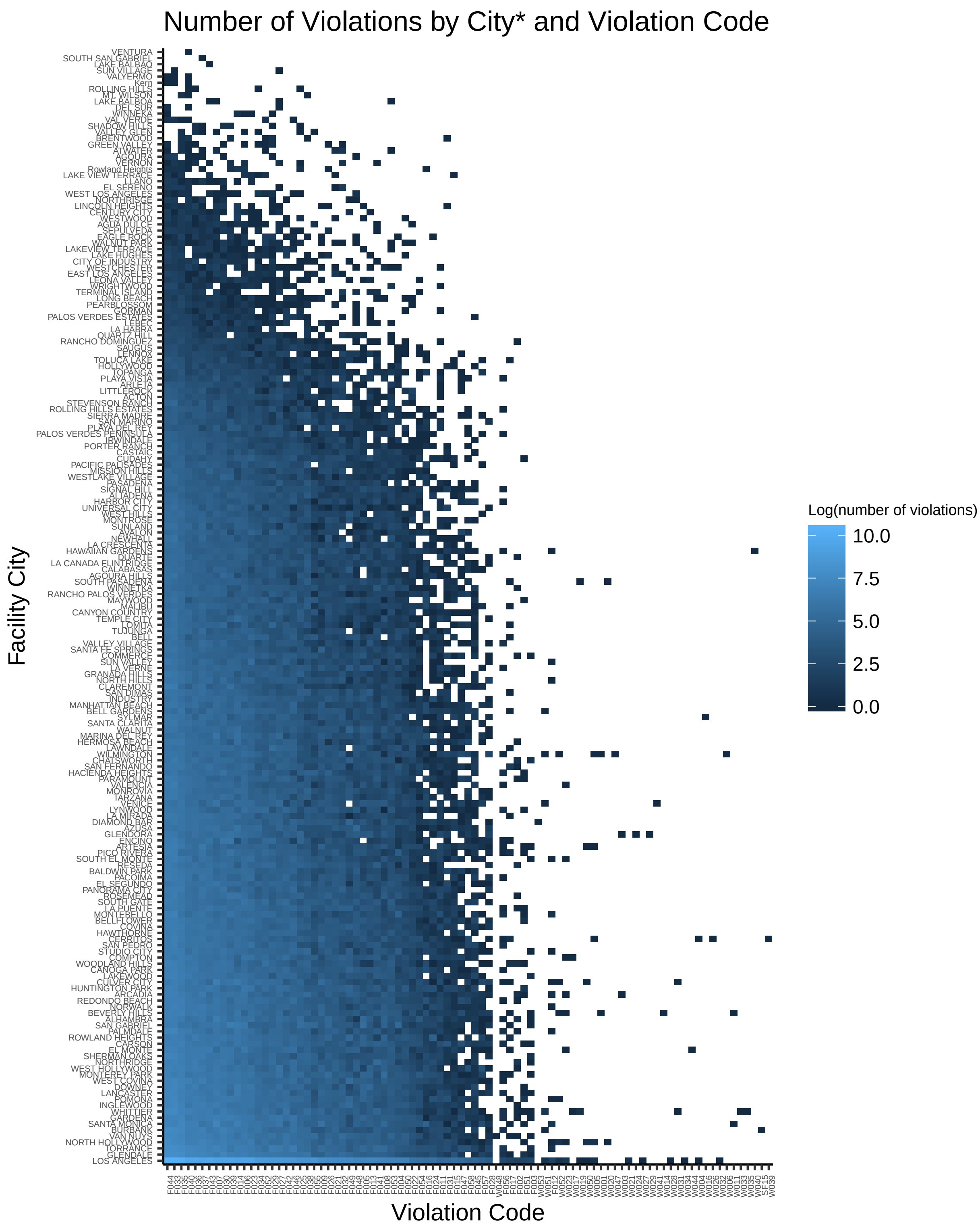

Figure 10: More than half of violation codes are sparsely represented (i.e. most are violation codes that begin with W or SF).

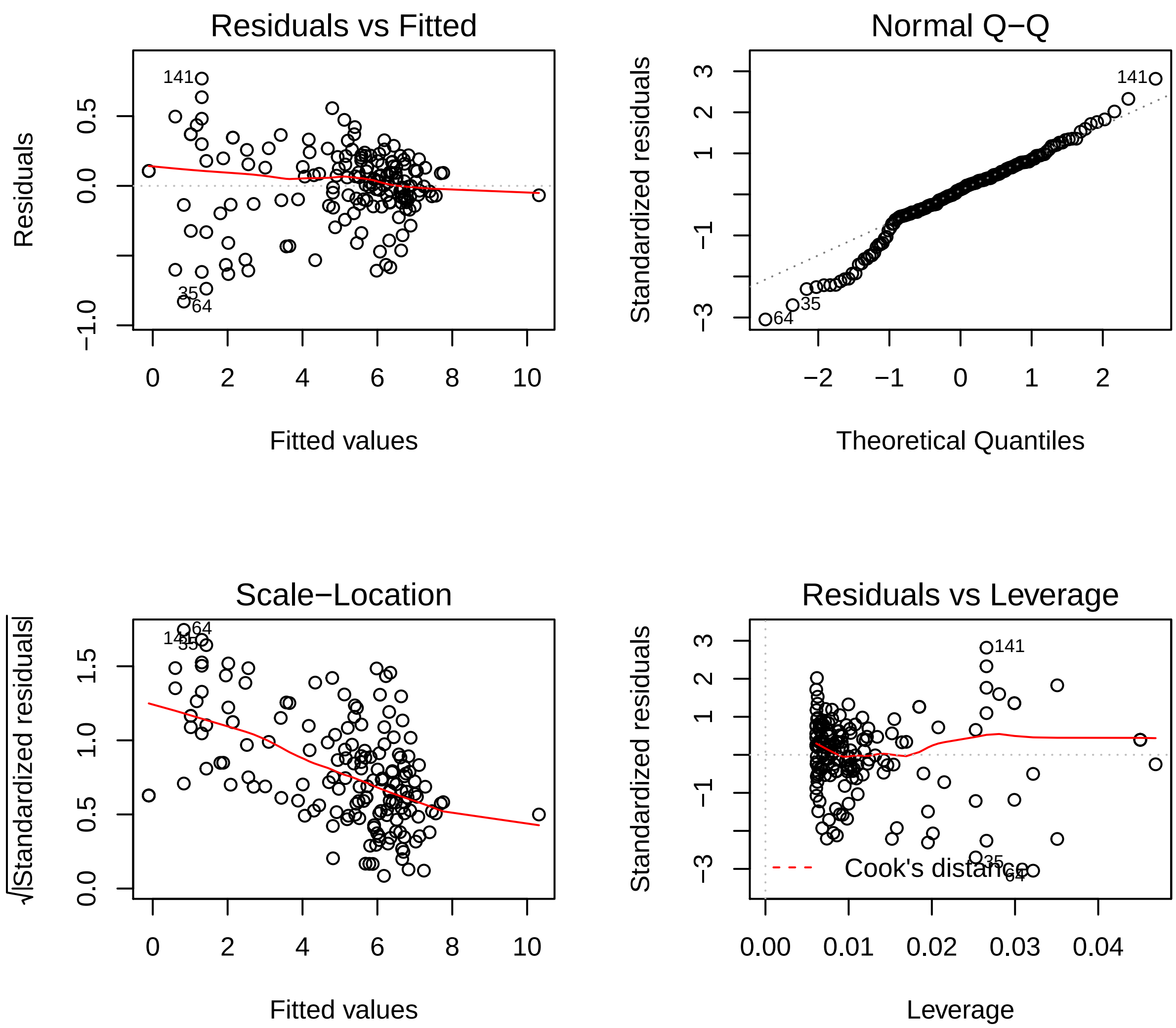

Figure 11: Model diagnostics for Santa Monica for health code F033.

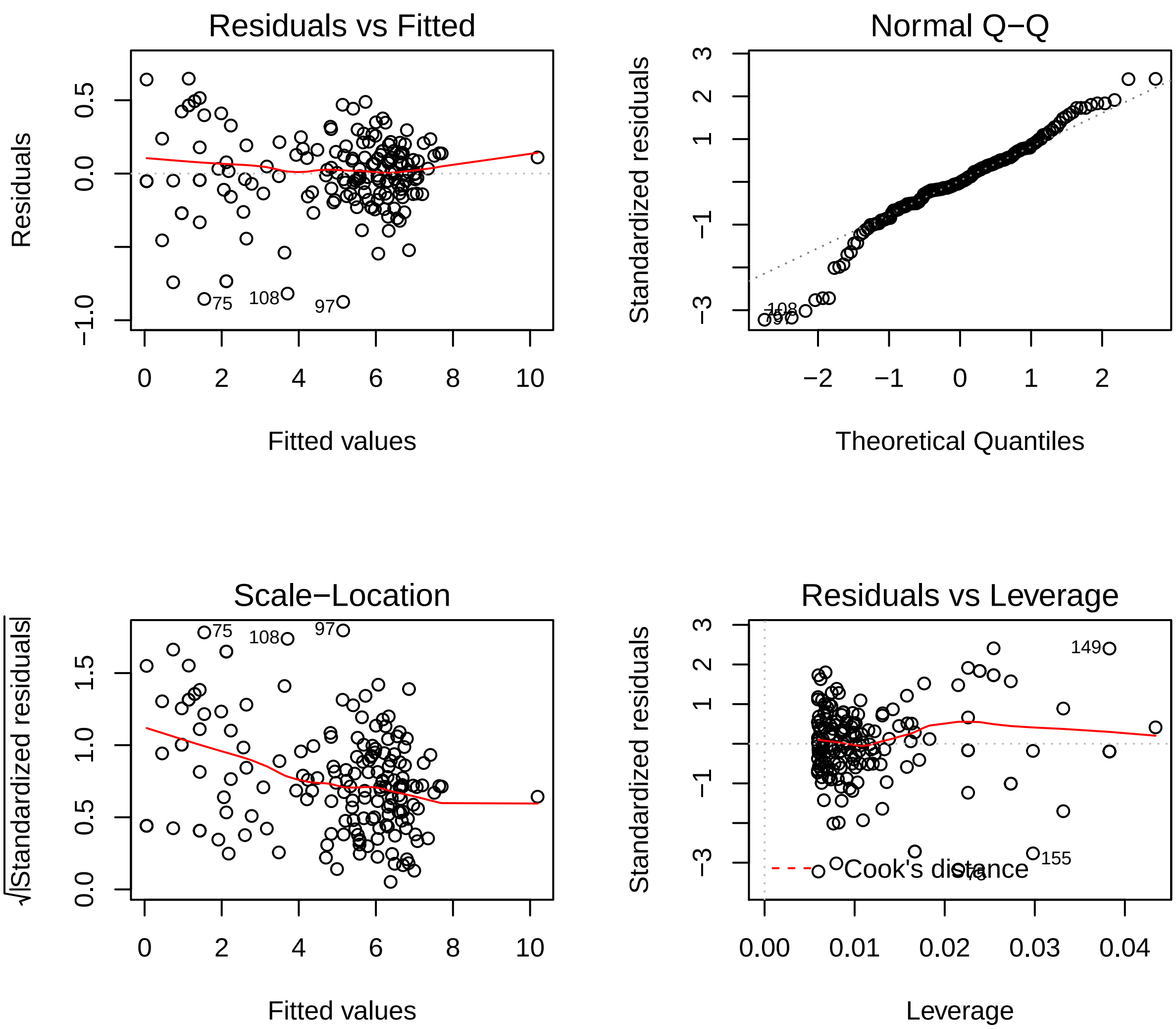

Figure 12: Model diagnostics for Santa Monica for health code F044.

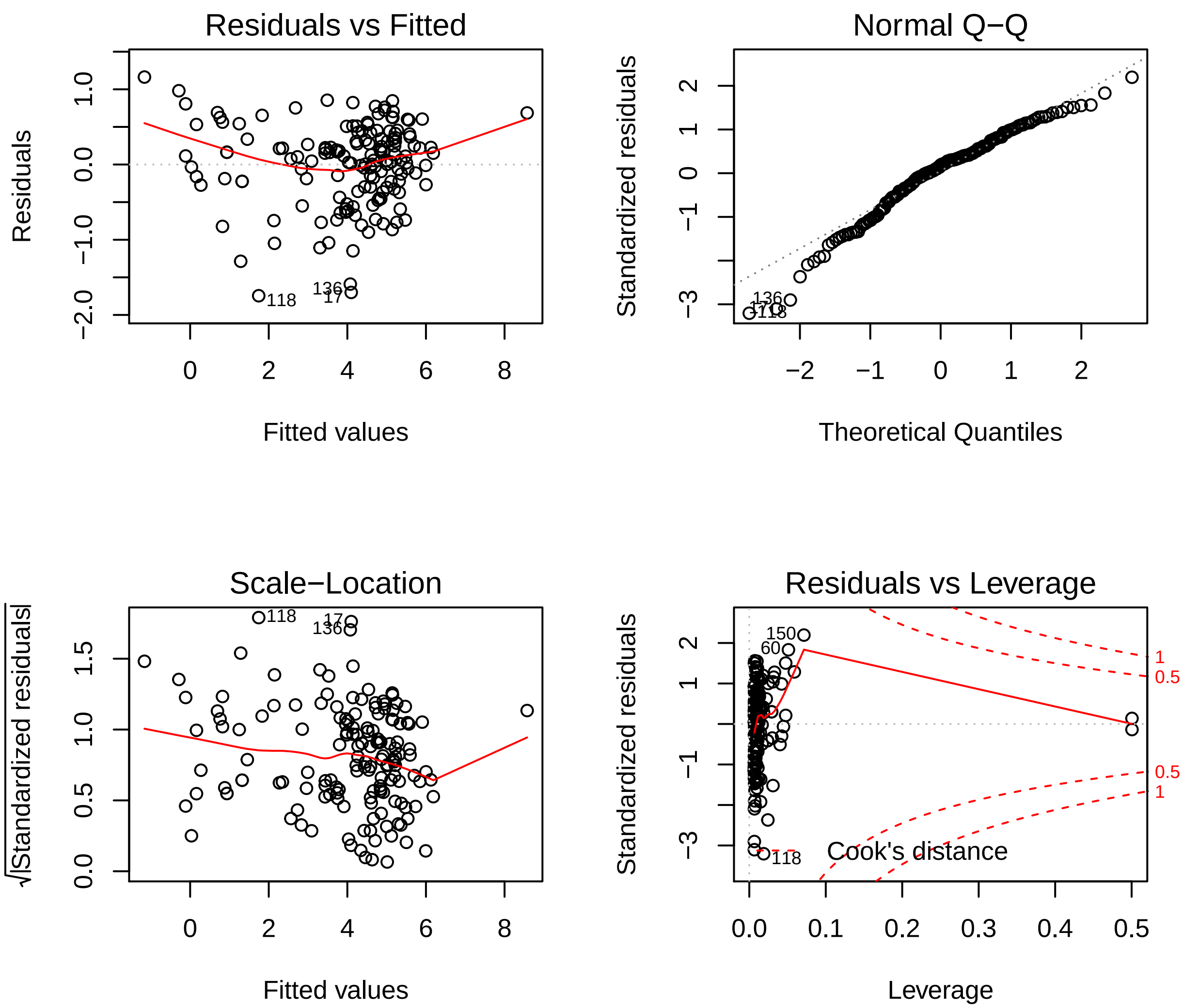

Figure 13: Model diagnostics for Avalon and Sun Valley for health code F023.

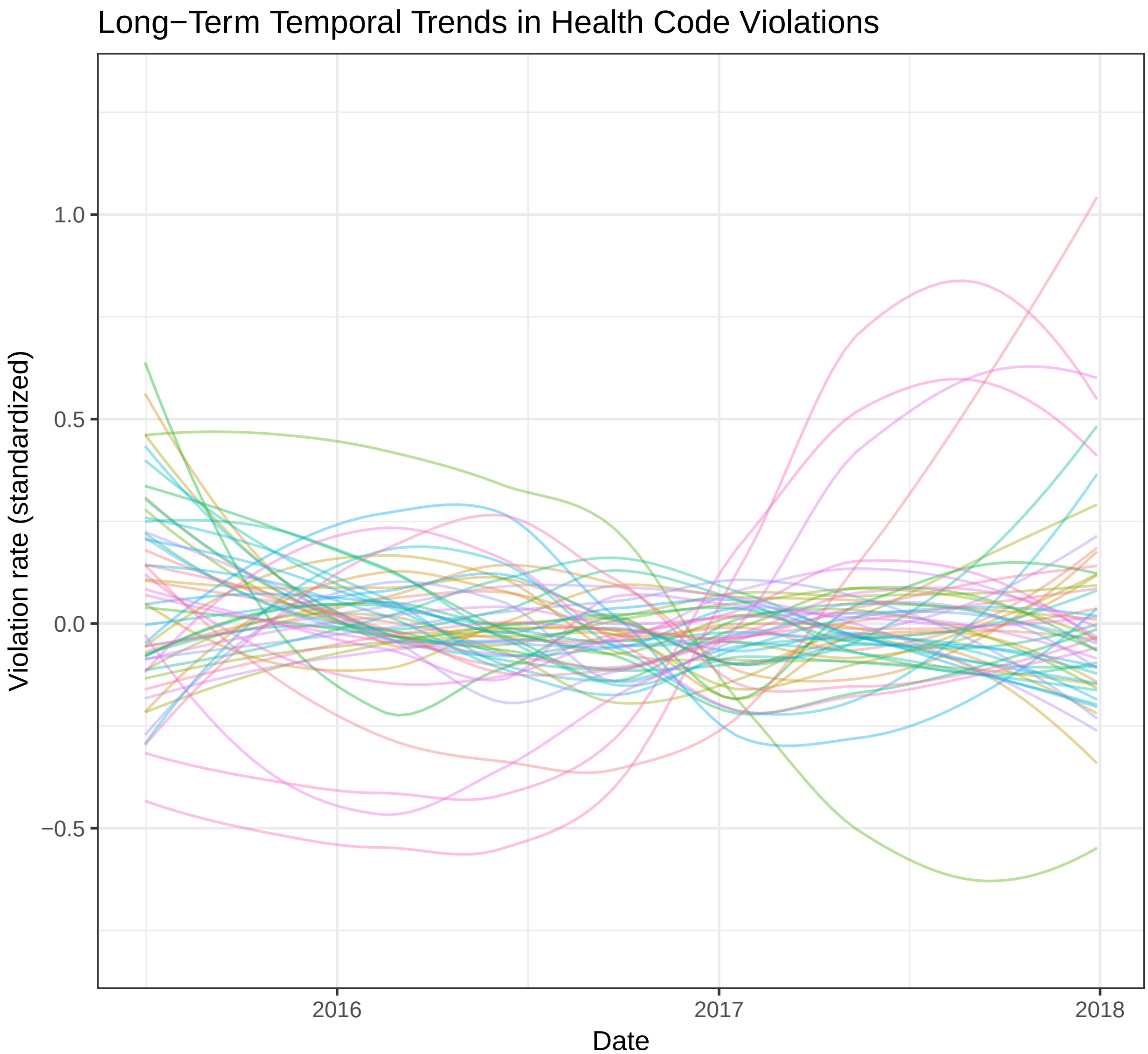

Figure 14: Each color corresponds to a different health code. A few health codes undergo substantial changes in the rate of recorded violations over time, while the rest are stable in comparison.