Our goal is to predict which restaurant inspections in the city of Chicago will result in critical violations. The longer a restaurant operates with an undetected violation, the greater the risk to the general public. To address this problem, we are examining data from three sources. As a beginning step in this work, we cleaned our main data sources, removing spurious entries and recoding variables numerically in ways that facilitated our analyses.

First, we are looking at the records of food inspection outcomes since 2010. This database contains all the food inspections done within Chicago over that period, which amounts to over 130,000 entries. For each inspection, we have the result: a pass, a pass with conditions, or a critical violation. Furthermore, we have access to the establishment name, the “risk category” of the facility, location, type (restaurant vs café, etc), as well as details on the inspection failure, if any.

We are combining the inspection data with weather records from the same period, obtained from National Oceanic and Atmospheric Administration (NOAA). These data were recorded at Chicago O’Hare international airport; we are using the daily high, low, and average temperature, the amount of precipitation, the amount of snowfall, and the average wind speed. From a previous analysis done by the City of Chicago, we believe the temperature data may predict inspection outcomes to some degree.

Finally, we are also examining 311 sanitation complaint records from the City of Chicago. We downloaded all the 311 sanitation complaints in the city, of which there were 112,000, or about 45/day. These complaints may reflect areas of the city where sanitation standards are reduced and therefore may predict where restaurants will fail their inspections.

We first examined whether the inspection outcomes depended on nearby sanitation complaints. To do this, we took each inspection record and found all the sanitation complaints that occurred within the previous week and within a small local distance (about a 3km radius). We then plotted a histogram of the number of complaints among those inspections with critical violations and compared it to a histogram for the inspections that passed or passed with conditions (Fig 1). We found that those inspections that passed had fewer recent local 311 complaints (histogram has more mass to the left; green line). Conversely, those inspections that failed were likely to have more sanitation complaints within the past week (red line). However, this effect is quite subtle; a classifier based on this information alone would be quite poor.

Next, we examined whether the inspection outcome depended on the weather. We plotted a histogram of the day’s high temperature for each inspection that critically failed (Fig 2A, panel A, red line). We did the same for those inspections that passed with conditions (blue line) and those that passed outright (green line). We found that the distribution of temperatures was skewed toward hot days for inspections that failed. That is, inspections were more likely to fail on hot days than on moderate ones. Interestingly, both very hot days and very cold days did not seem to affect inspection outcome; only above-freezing to 70° days were more likely to pass while 80° to 90° days were more likely to fail. However, as with the number of nearby sanitation complaints, this effect is very subtle.

The daily low was even less predictive than the daily high (Fig 2, panel B). Similarly, the day’s wind speed and precipitation had not predictive power at all.

In terms of the inspection data itself, we examined potential relationships between inspection outcome and temporal variables such as month or day of inspection as well as changes in inspection outcome for different types of facility, zip code, and risk level. The results of the inspection varied by risk category when examining “low risk” institutions. Such vendors were less likely to receive a “pass with conditions” rating and more likely to fail outright than their peers in the other two risk categories (Figure 3). Zip code of establishments also revealed some interesting patterns, with some zip codes and neighborhoods more likely to assign “pass with conditions” than others (Figure 4). This should definitely play a role in our analysis of this as a three-way classification problem, as these results were lumped into “pass” in the binary classification the city performed. The “facility type” list has over 400 different facility types listed and a simpler way of categorizing these establishments is needed. We will devise such a method prior to our next milestone.

Our temporal variables yielded little insight, as every month had a relatively identical ration of pass / pass with conditions / fail (Figure 5). Inspections of high-risk establishments were carried out at the same rate relative to lower-risk establishments throughout the year, indicating any month-to-month variation was not an effort to rapidly assess high-risk vendors as quickly as possible in a given year. It does appear as if more inspections occur during spring and fall months (Figure 6), and this could relate in interesting ways to our temperature and weather findings.







