A for Effort? Using the Crowd to Identify Moral Hazard in NYC Restaurant Hygiene Inspections

Jorge Mejia
Kelley School of Business, Indiana University, Bloomington, Indiana 47405, jmmejia@iu.edu

Shawn Mankad
Samuel Curtis Johnson Graduate School of Management, Cornell University, Ithaca, NY 14853, spm263@cornell.edu

Anand Gopal
Robert H. Smith School of Business, University of Maryland, College Park MD 20742, agopal@rhsmith.umd.edu

Online Appendix

Appendix A: New York City Restaurant Inspections Program Details

There are several types of violations as specified in the in the NYC Health Code (New York City Department of Health and Mental Hygiene 2016a). First, violations are classified as "critical" or "general". Critical violations are ones that contribute the most to foodborne illness and pose a significant risk to consumers, thereby receiving more points than general violations. For example, failing to maintain certain ingredients at a safe temperature is a critical violation and receives between 7 and 28 points, while failing to maintain a clean toilet facility is a general violation and receives between 2 and 5 points. Furthermore, critical violations that pose an immediate "public health hazard" receive the highest number of points. If a restaurant does not correct such violations before the end of the inspection by implementing the required processes, it may result in the restaurant being closed.

The number of points received for a particular violation also depends on the condition level, which is the extent and frequency of the violation. Some violations have more condition levels and parameters than others. Conditions can vary from level 1 to level 5, with higher levels receiving higher numbers of points and signifying a more severe violation (New York City Department of Health and Mental Hygiene 2016c). For example, the presence of a single contaminated food item would constitute a lower condition

level (level 1 and 7 points), whereas the presence of 4 or more different contaminated food items would earn a higher condition level for the violation (level 5 and 28 points).[1]

At the end of each inspection, the inspector reviews the results with the operator, explains the violation and condition scores and makes suggestions to improve food safety. The inspector then issues an inspection report, which contains a list of all violations and their corresponding points and severity, and the total inspection score (New York City Department of Health and Mental Hygiene 2016a). Depending on the specific violations identified, restaurants are required to financial pay fines, which range from $200 to $2000, and may be higher for repeated violations (New York City Department of Health and Mental Hygiene 2016a). As a further penalty, restaurants are automatically closed if they score a grade of C in three consecutive inspection cycles (New York City Department of Health and Mental Hygiene 2016b).

Figure A1. NYC Restaurant Hygiene Cards

(New York City Department of Health and Mental Hygiene 2016d)

Table A1. Cycle 3 Inspection Scores for Restaurants Achieving an A in Cycle 2

		Cycle 3 Grade			
		A	PA	PB/C	Total
Cycle 2 Grade	A	85.51%	10.07%	4.42%	100.00%
	PA	2.86%	90.44%	6.70%	100.00%

[1] A full description of the critical and general violations can be found here: https://www1.nyc.gov/assets/doh/downloads/pdf/rii/self-inspection-worksheet.pdf

Appendix B: Estimating the Impact of a "P" Grade on Hygiene Ratings in Re-Inspection

Restaurants that do not clear the bar for an A grade on initial inspection receive a P grade, indicating that their final grade within that inspection cycle is pending. Since the hygiene grade is an important signal in the restaurant industry (Jin and Leslie 2009), we would expect a significant improvement in the re-inspection score within the same inspection cycle. Recent reports of inspections program in NYC show this to be the case (New York City Department of Health and Mental Hygiene), confirming the notion that restaurants do respond proactively to the initial inspection. We also estimate this statistically to control for other co-varying factors. To estimate the effect of receiving a P grade on the improvement observed in the same restaurant's re-inspection score, we employ a longitudinal difference-in-differences (DiD) model, which allows the same restaurants to serve as treatment and control at different points in time, and is thus an appropriate methodology for causal inference (Bertrand et al. 2002). The idea behind the DiD model is to examine a group of treated units before and after the treatment. In our case, restaurants are considered part of the *treatment* group when an inspection occurs that has a possibility of a P grade (initial inspection) and are considered part of the *control* group when there is no possibility of a P grade (re-inspection). The control group is an important part of the DiD framework since other factors that influence hygiene might have changed over time. Additionally, the DiD model removes any time-invariant restaurant-specific heterogeneity that may influence the outcome (see Lechner (2011) for a review of this procedure).

The DiD framework is particularly appropriate for our context for several reasons. First, the treatment (initial inspection) is randomly assigned and all restaurants receive the treatment at some point in the dataset. This is of particular importance to identify a treatment effect in longitudinal models (Athey and Imbens 2006). Second, we check on the common trends assumption in the data (i.e. that the differences in the expected control outcomes over time are not related to being part of the treated or control group in the post-treatment period). The implication is that if the treated group had not been subjected to the treatment, it would have experienced the same time trends. Intuitively, the common trends assumption is unlikely to be violated in our context, since all groups in our data belong to a single category of merchants,

restaurants, and are located in the same geographical region. Nevertheless, we test this assumption below. Third, having a high number of time periods (particularly similar time periods) and groups (particularly similar restaurants) of control units is important, as it has been shown to provide more precise estimation of treatment effects, more reliable testing of the common trends assumption, and more precise inference (Lechner 2011). However, one limitation of our DiD approach is that while restaurant inspection dates are random, the re-inspection (in the case of a P-grade) typically occurs between 2 and 4 months. Thus, the restaurant may anticipate the re-inspection visit, which could potentially violate the strict exogeneity assumption of DiD. To check for the appropriateness of the DiD framework, we test for the common trends assumption using the leads and lags methods of Autor (2003) and Angrist and Pischke (2009). We find no evidence of violations of the common trends or the exogeneity assumptions.

As before, for the DiD model, the unit of analysis is restaurant-time period, where the unit of time is one *month*. The outcome variable is the numerical inspection score, available from the inspection report. We include two types of controls, as suggested by Imbens and Wooldridge (Imbens and Wooldridge 2007): group-level and treatment-level. Group-level controls, which improve identification in DiD models, include all observable restaurant characteristics, such as restaurant price point, location, and overall Yelp star rating. Treatment-level controls, which help account for within-group variation and reduce standard errors (Imbens and Wooldridge 2009), include inspector characteristics. Summary statistics for the final dataset used in the DiD model are shown in Table A2. The estimated DiD model is of the form:

$$INS_{it} = \beta_0 + \beta_1 TreatmentGroup_{it} + \beta_2 AfterTreatment_{it} + \beta_3 TreatmentGroup_{it} * AfterTreatment_{it} + \boldsymbol{\gamma Controls_{it}} + \varepsilon_{it}, \quad (7)$$

where INS_{it} is the numerical inspection score for restaurant i during time period t. Table A3 displays the results from this model, which show a highly negative and significant (-6.25; $p<0.001$) average treatment effect (ATE) of an initial inspection on inspection scores. This indicates that hygiene scores improve considerably for the average restaurant between the first and second inspection within the same inspection cycle. Moreover, we see a differential ATE for restaurants in different price segments, with high-priced restaurants (*3-4 Yelp dollar signs*) having less negative ATE, as seen by the positive and significant

interaction coefficient (1.81; $p<0.01$)[2]. Higher priced restaurants therefore appear to be more stable in their hygiene performance before and after initial inspections, compared with restaurants in lower price ranges.

Table A2. Summary Statistics for Difference-in-Differences Model

Measure	**Value**
Number of restaurants	21,488
Mean inspections per restaurant	6.33
Mean violation score per inspection	15.11
Restaurant characteristics:	
Mean Price point	1.84
Mean Yelp Rating	3.54
Mean Number of Reviews	102

Table A3. Effect of Initial Inspection on Hygiene Inspections Scores Using Difference-in-Differences Model, DV= Inspection Scores

Variable	**Estimate (SE)**
Treat (Initial Inspection)	-6.25 (1.51) ***
Rating	-5.22 (2.60) *
Reviews	0.00 (0.00) ψ
Pricepoint	0.01 (0.00) ψ
Treat*Higher Price ($$$-$$$$)	1.81 (0.59) **
Group and Treatment Controls	Included
Groups	21,488
Observations	136,503
R^2	0.55
Robust standard errors in parentheses Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 'ψ' 0.1	

[2] The baseline was the low-priced restaurants *(1-2 Yelp dollar signs)*.

Appendix C: Sensitivity Analysis and External Validation of the SMASH Dictionary

In this section, we perform multiple validation studies of the SMASH dictionary, starting with the use of the Naïve Bayes classifier to select initial seed words, followed by an external validation of the final dictionary using a dataset from a different city and time interval.

As mentioned in the main text, in principle several classification or supervised learning methods can be utilized to learn words that are associated with poor hygiene from the labeled reviews. As such, we perform a benchmarking study here to compare goodness of fit of three of the most popular supervised learning techniques: Naïve Bayes, Support Vector Machines, and Logistic Regression with Lasso (ℓ_1 penalty). The benchmarking is performed using 1000 repetitions of cross-validation, where the model is trained on a subset of the labeled data and tested on the remaining held-out subset. The training data is constructed by randomly sampling without replacement 70% of the original data, with the remaining 30% creating the test set. Results are provided in Table A4 showing that (i) the supervised learning methods are all in general accurate, achieving over 92% accuracy (or equivalently 8% error rate); (ii) there are no significant differences between the classifiers on our data, though technically Naïve Bayes has the best accuracy by a negligible amount. Moreover, though not shown, we find significant overlap in the initial seed words across different supervised learning methods, adding evidence that the dictionary and subsequent analysis results are not sensitive to the choice of classifier.

As an additional validation check on the final SMASH dictionary, we employ an independent dataset from a recent competition sponsored by Harvard, Yelp.com, the City of Boston, and DrivenData to use the text of online reviews to predict inspection scores for restaurants in Boston (DrivenData 2016). The Boston restaurant inspection program, as in the NYC RIP, randomly inspects restaurants for hygiene-related issues. As part of the crowd-sourced competition, submissions were elicited from the public for two months. The winner was the submission with the lowest prediction Root Mean Squared Logarithmic Error (RMSLE), and the top three teams were awarded financial prizes. To assess the performance of our methodology, we use the program's publicly available training and test datasets, which include inspection scores from the inspection agency in Boston and restaurant and review characteristics from Yelp. We

compare the predictive performance of a model using the SMASH word counts with the winning algorithm from the competition on the Boston dataset (see Table A5). We find that a regression model based on the SMASH scores would have placed in the top five, indicating similar predictive performance to the winning entries. We emphasize that all competing models in Table A5 were trained on data from Boston, where our dictionary *was not retrained on the competition dataset.* Recall that the SMASH dictionary was constructed on NYC reviews collected over different time periods from the competition data. These results demonstrate out-of-sample validation for the SMASH dictionary and its generalizability to other urban areas.

Table A4. Misclassification Error Rates for Supervised Learning Methods to Seed the Dictionary Creation Procedure

Method	**Misclassification Error Rate**
Naïve Bayes	7.814 (0.037)
Support Vector Machines	7.943 (0.036)
Logistic Regression with Lasso	7.856 (0.037)

Table A5. External, Out of Sample Validation for SMASH using the DrivenData Competition.

Rank	**User or team**	**Current evaluation score (Weighted RMSLE)**	**Submission timestamp**
1	LilianaMedina	0.8901	July 6, 2015, 3:07 p.m.
2	qwang	0.8931	July 6, 2015, 8:10 p.m.
3	singhs	0.9113	July 7, 2015, 2:03 p.m.
4	furiouseskimo	0.9179	July 5, 2015, 6:27 p.m.
NA	Using SMASH Dictionary Approach	0.9221	N/A
5	devonb	0.9428	July 7, 2015, 10:37 p.m.
6	Moules-frites	0.947	July 6, 2015, 6:38 p.m.
7	windows	0.9547	July 6, 2015, 12:47 a.m.
8	fizzicks	0.9547	July 7, 2015, 9:51 p.m.
9	scottkam	0.9635	July 5, 2015, 8:39 a.m.
10	micah	0.976	July 7, 2015, 11:37 p.m.
…	…	…	…
24	Kathrynjbarger	1.8499	July 7, 2015, 9:11 p.m.

Appendix D: Ridge, Lasso, Elastic Net, and Random Forests Analyses

In the main paper, we adopt an analysis-of-variance approach to estimate the correlation between the SMASH word counts and the hygiene inspection score. In this appendix, we consider approaches that are based on improving predictions per se, rather than the linear regression approach, in order to establish the robustness of the SMASH dictionary. We implement and test several statistical learning techniques to examine the focal relationship between SMASH word counts and hygiene scores. We first perform *regularized linear regression* to complement the linear model-based results presented in the main paper. The idea behind regularized linear regression is to constrain the complexity of the problem by shrinking some coefficients towards or to zero (Hastie et al. 2009). These methods can significantly reduce the variance of the estimated coefficients and help with variable selection (Seber and Lee 2003) with a view to enhancing the predictive power of the analysis. The three basic classes of regularized regression we consider here are *ridge regression* (Hoerl and Kennard 2000), *LASSO* (least absolute shrinkage and selection operator) (Tibshirani 1996) and *elastic nets* (Zou and Hastie 2005). In the case of ridge regression, the regularization is the squared l_2 norm, whereas in the LASSO, the penalty is the l_1 norm, while in elastic nets the penalty is a convex combination of the LASSO and ridge penalties.

$$min_\beta \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_0 x_{i,j}\right)^2 + \lambda_1 \sum_{j=1}^{p}\beta_j^2 \qquad \text{(Eq. 1: Ridge)}$$

$$min_\beta \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_0 x_{i,j}\right)^2 + \lambda_2 \sum_{j=1}^{p}|\beta_j| \qquad \text{(Eq. 2: Lasso)}$$

$$min_\beta \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_0 x_{i,j}\right)^2 + \lambda_1 \sum_{j=1}^{p}\beta_j^2 + \lambda_2 \sum_{j=1}^{p}|\beta_j| \qquad \text{(Eq. 3: Elastic Net)}$$

where for subject i, y_i is the value of the dependent variable, $x_{i,j}$ is the value of predictor variable j, $j = 1, \ldots, p$, and λ_1 and λ_2 are tuning parameters that modulate the importance of fit to the observed data versus parsimony of the resulting model (i.e., the degree of coefficient shrinkage).

We applied each of three methods to variables described in Table 1 of the main paper with the same specification presented in the main text (used to generate the results in Table 4 of the main paper), including all observable restaurant and reviewer characteristics in the baseline set of predictor variables. As

recommended by Hastie et al. (2009), we randomly split our data into equally-sized training and test datasets. Within the training dataset, we perform 10-fold cross-validation to choose the values of λ_1 and λ_2 that minimize the cross-validation mean squared error (MSE). Figure A2 plots the coefficients paths (left panels) and training set MSE for the ridge, LASSO, and elastic net models (right panels). The SMASH word counts are the second chosen variables for LASSO and elastic nets only after the rating.

The coefficients for the selected set of variables resulting from each method are shown in Table A6. It is worth noting that the same set of independent variables is selected, regardless of method used. The estimates for our main variable of interest, the SMASH word counts (WC), as well as Ratings and Reviews, are highly consistent with the coefficients results reported in the main paper. Interestingly, the variable selection aspect applied to the categorical variables of borough and cuisine resulted in a single borough (Brooklyn) and cuisine (pizza) being selected. We note that these were also included as controls in the models reported in the main paper, along with all other boroughs and cuisines. Comparing the test set MSE of the linear model presented in the main paper (Table 4) with the models selected here using regularized regression, as reported in Table A7, we find that OLS provides a comparable MSE. We thus find confirmation of the results reported in the main paper using regularized linear regression techniques as well, indicating that the SMASH wordcounts and their correlations with hygiene inspection scores are robust to alternative methods as well.

While the use of regularized linear models provides significant advantages over other statistical techniques in terms of simplicity and interpretability, there is a risk that we may miss significant non-linear effects and interactions prevalent in the data.[3] We therefore also implement random forests to consider these non-linear and interactive effects on the results observed in the main paper (Breiman 2001b). Following Hastie et al. (2009), we use nominal values for the number of trees to grow (i.e., 500) and establish a minimum node size (i.e., 5). Moreover, we choose the number of variable candidates to sample at each split through 10-fold cross validation on the training dataset. Figure A3 plots the number of trees versus the

[3] We thank an anonymous reviewer for pushing us to consider non-linear models in addition to OLS regression.

training set MSE while Figure A4 displays the variable importance plot (Breiman 2001a), which is used to examine the importance of the top variables selected within the model. As is evident, the random forest selects the rating, time period, SMASH word counts and price as the most important variables. Also, while the test set MSE for the random forest is lower than ridge, LASSO, and elastic net models, the linear model presented in the main paper still achieves slightly lower MSE (Table A7). These results again show that the estimated results in the main paper are robust, as is the SMASH dictionary. The models presented here were implemented using language R version 3.3.1 (R Core Team 2017) using packages glmnet (Simon et al. 2011) and randomForest (Liaw and Wiener 2002).

Figure A2. Coefficients Path and MSE for Ridge, Lasso, and Elastic Net

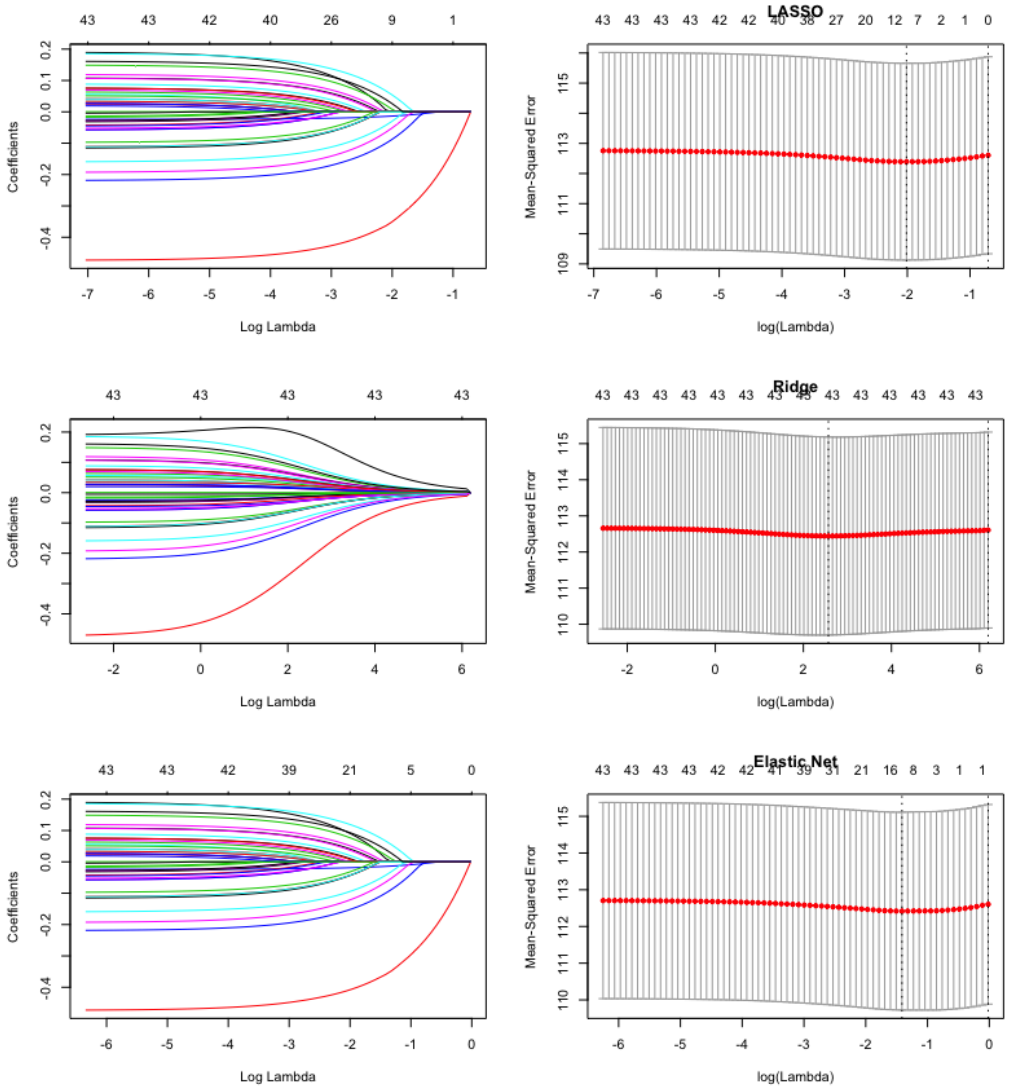


Figure A3. Trees Vs. MSE for the Random Forest

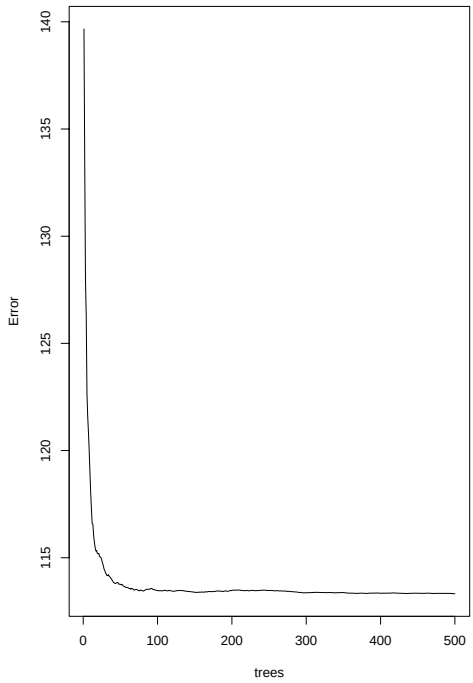


Figure A4. Variable Importance Plot for Random Forest

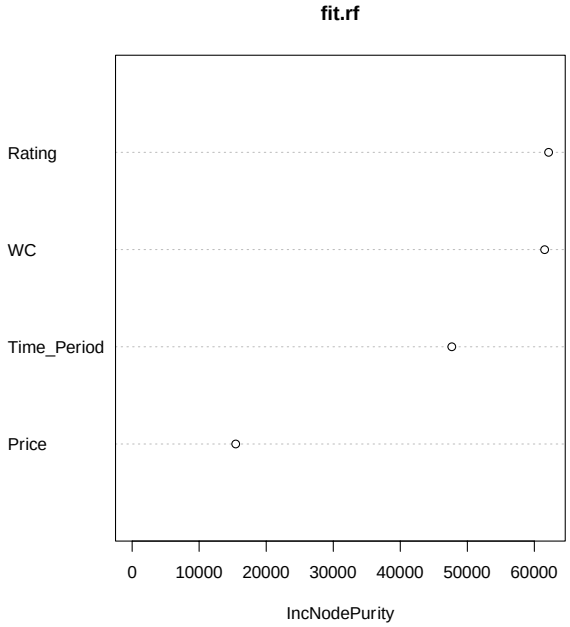


Table A6. Model Coefficients for Ridge, Lasso, and Elastic Net Regressions

Model	**Ridge**	**Lasso**	**Elastic Net**
Variables	**Coefficients**		
Intercept	20.62	20.64	20.64
WC (WordCount)	0.49	0.48	0.48
Rating	-0.47	-0.47	-0.47
Reviews	0.00	0.00	0.00
Time_Period	-0.02	-0.02	-0.02
Price	-0.05	-0.05	-0.05
Borough_Brooklyn	0.11	0.11	0.11
Cuisine_Pizza	0.10	0.10	0.10
Others	.	.	.

Table A7. Test set MSE of the linear model in the paper with the regularized regression and random forest models

Model	**MSE**
OLS (Table 3, Model 1 in main paper)	108.8
Ridge	112.49
LASSO	112.64
Elastic Net	112.7
Random Forest	111.50

Appendix E: Investigating Serial Correlation in SMASH Term Usage

Previous literature (Li and Hitt 2008, Muchnik et al. 2013) has discussed the observed herding effect in online reviews, i.e., situations where a reviewer is influenced by, and thus generates online review ratings, correlated with reviews for the same restaurant (or product) posted previously by other consumers. It is not clear whether this herding behavior occurs at the level of the individual phrases present in the SMASH dictionary. Aggregated at the level of the restaurant within a time-period, we can account for the effects of serial correlation – we do so in the analysis provided in the main paper at the month level by lagging independent variables and modeling an AR-1 autoregressive process. The results from this analysis show a consistent and clear correlation between SMASH word counts and hygiene inspection scores.

Another possible source of herding behavior could emerge from consumers using specific words or phrases in their reviews, as a result of their being used in previously written reviews for the same restaurant by other consumers. For instance, if one review mentions the word "*hair*" in a review, it is possible that the subsequent review (or reviews) also include the word "*hair*" in their text. If such a pattern exists, the resulting SMASH word count may correspond to herding rather than actual observations of poor hygiene. To examine this issue statistically, we test for potential auto-correlation of each SMASH dictionary term at the individual review level using the Ljung-Box test (Ljung and Box 1978). Specifically, for every restaurant and SMASH term combination, we form a time-series of word counts ordered by the date of the review posting. As shown in Table A8, there are over 100,000 individual time-series for each word-restaurant combination, wherein we test for potential serial correlation of up to lag order 5 using a 5% significance level. We also repeat the exercise to test for serial correlation of up to lag order 5 at the daily-level of aggregation, rather than at the level of an individual review.

The results show that for approximately 4.5% of the time-series at both the individual review and daily levels, we observe statistically significant serial correlation. In other words, in 95.5% of the sample constructed, we find no evidence of serial correlation in the way SMASH terms are used within online reviews. Moreover, given the large number of hypothesis tests being conducted, we would expect erroneous

rejection of the null hypothesis 5% of the time – matching the significance level (Benjamini and Hochberg 1995). Using the procedure outlined in Benjamini and Hochberg (1995) to control for the false discovery rate (i.e., minimize false positives when conducting so many thousands of hypothesis tests), we find less than 2.3% of the 103,917 time-series exhibit signs of serial correlation. Thus, based on the presented results, we conclude that serial correlation of usage of specific words across time is unlikely to be a significant issue in our data.

Table A8. Serial Correlation Box-Ljung Test Results for Individual SMASH term usage

Resolution	**# Smash Terms**	**# Restaurants**	**# Hypothesis Tests**	**% Reject H_0**	**% Reject H_0 using FDR[4]**
Review-level	142	6490	103,917	4.41	2.20
Daily-level	142	6490	103,917	4.14	1.99

[4] "FDR" refers to "False Discovery Rate" procedure (Benjamini and Hochberg 1995). A 5% significance level is used and rejecting the null hypothesis is evidence for serial correlation. The number of hypothesis tests does not equal number of terms multiplied by number of restaurants - not all terms are used in reviews of every restaurant.

Appendix F: SMASH Dictionary Comparison

In this appendix section we compare our SMASH dictionary with that of other works, namely Kang et al. (2013) and Schomberg et al. (2016). Table A10 shows a subset of the SMASH dictionary along with the word lists published with both papers.[5] Illustrating a common component amongst all three dictionaries, there is some degree of overlap in the keywords. For example, the terms common to both SMASH and Kang et al. (2013) are "dirty" and "gross", while the terms common to both SMASH and Schomberg et al. (2016). are "awful", "dirty", "hell", "horrible", "roach", "stomach", "sucks", "filth" – all keywords that signal potentially poor hygiene conditions. Consequently, as shown in Table A9, the dictionaries result in word counts that are positively correlated. However, the relatively moderate levels of association also indicate that the dictionaries capture different types of information.

Note, that the Schomberg et al. (2016) word-list is compiled specifically to capture food-borne diseases and was manually picked by the researchers, based on their own sense for what is important. For instance, the Schomberg et al. dictionary contains terms seemingly focused on illness and disease, such as "hospital", "vomiting", "food poisoning", and "diarrhea". In contrast, our dictionary is generated from the text in reviews in a relatively unsupervised manner. Once the seed list of words is generated using the Naïve Bayes method, we augment these by adding synonyms, but for the most part, the words in the dictionary are based on the text in reviews and broadly associated with hygiene rather than diseases per se. The value of our expanded approach to identifying words can be viewed best through examples. Our dictionary includes the phrases "wife found", "crawling around inside", "food [is] disgusting", "moth-eaten" as keywords, which appear to be indicative of some hygiene-related issue in the actual review text. Such phrases do not appear on the manually created wordlist such as Schomberg et al. (2016). Thus, we believe our dictionary captures hygiene-related information provided by consumers potentially in ways that are not easily captured by manually selected keywords by experts.

[5] Unfortunately, the Kang et al. (2013) conference proceeding does not provide specific keywords and therefore a full comparison is not possible.

Nevertheless, to more fully compare the performance of the SMASH dictionary with Schomberg et al. (2016), we re-estimate the linear models reported the main paper and the alternative methodologies explored in Appendix D. Specifically, we replaced the SMASH dictionary word counts in the models for the Schomberg et al. (2016) word counts in our specifications. In the linear model, we find that the Schomberg et al. (2016) word counts have a positive and statistically significant (at the 0.05 level) effect on the next hygiene score of the restaurant. However, we also find that replacing the Schomberg et al. (2016) word counts with the SMASH word counts results in a 5-7% improvement in the out of sample MSE, as shown in Table A11.

Additionally, to see which variable is more important to predictive accuracy within a non-parametric setting, we estimate a random forests model (Breiman 2001b; see in Appendix D) including both the SMASH word counts and the Schomberg et al. (2016) word counts. We used typical values for the number of trees to grow (i.e., 500) and establish a minimum node size of 5. Moreover, we choose the number of variable candidates to sample at each split through 10-fold cross validation on the training dataset. We find that the SMASH word counts have a higher Node Purity in the variable importance plot: The SMASH word counts have a node purity of 154662.3 while the Schomberg et al. (2016) word counts have a node purity of 136657.9. In other words, the random forests results show that both the SMASH and Schomberg et al. (2016) word counts are useful for prediction, though the proposed SMASH counts is a more powerful predictor.

Thus, consistent with the results presented above, we conclude that the SMASH dictionary provides significant benefits in the prediction of hygiene scores from what had been established in the literature.

Table A9. Correlation at the review level between word counts using different dictionaries

	SMASH	**Kang et al. (2013)**	**Schomberg et al. (2016)**
SMASH	1.000	0.192	0.276
Kang et al. (2013)	0.192	1.000	0.406
Schomberg et al. (2016)	0.276	0.406	1.000

Table A10. Subset of keywords contained in the SMASH, Kang et al. (2013), and Schomberg et al. (2016) dictionaries. Note that only a partial list is available for Kang et al. (2013) and the SMASH dictionary should be applied after removing stop words

SMASH (Subset of terms)	Kang et al. (2013)	Schomberg et al. (2016)
Bugs, cockroaches, contaminate, contaminating, defecate, dirty disgusting food, eaten saw cockroaches, eggs gross, filthy, food disgusting, food horrible, gross, hair baked eggs, health department, horrendous, ill-scented, inedible, inhuman, insect bite, insects, insects food, like insects, moth-eaten, wiping nose	dirty, gross, mess, restroom, smell, sticky	ache, adorable, affordable, asshole, awful, bathroom, bitch, burnt, Christ, cigarette, clean, craving, delicious, diarrhea, DIRTY, dirty, dishes, employees, excellent, exclaim, fabulous, favorite, filthy, fish, food poisoning, fuck, hell, high quality, horrible, hospital, Humid, I found a, I love, Jesus, Mice, microwaved, Mold, nausea, oldschool, Pain, Pool, professional, puke, pushy, recommend, roach, rotten, septic, service, Shit, Sick, spider, stench, stomach, sucks, terrible, the best, toilet, truck, vomiting, wash hands

Table A11. Comparing test set MSE using SMASH word counts with test set MSE using Schomberg word counts

Model	MSE (SMASH WC)	MSE (Schomberg et al. (2018) WC)	% Difference
OLS (Table 3, Model 1 in main paper)	108.8	116.4	7%
Ridge	112.49	119.2	6%
LASSO	112.64	120.5	7%
Elastic Net	112.7	120.6	7%
Random Forest	111.50	117.0	5%

Appendix G: Examining SMASH Trends Over Time

Figure A5. Comparable to Figure 3 in the main paper but using polynomials shows similar downward and upward trends after the initial inspection and re-inspection, respectively

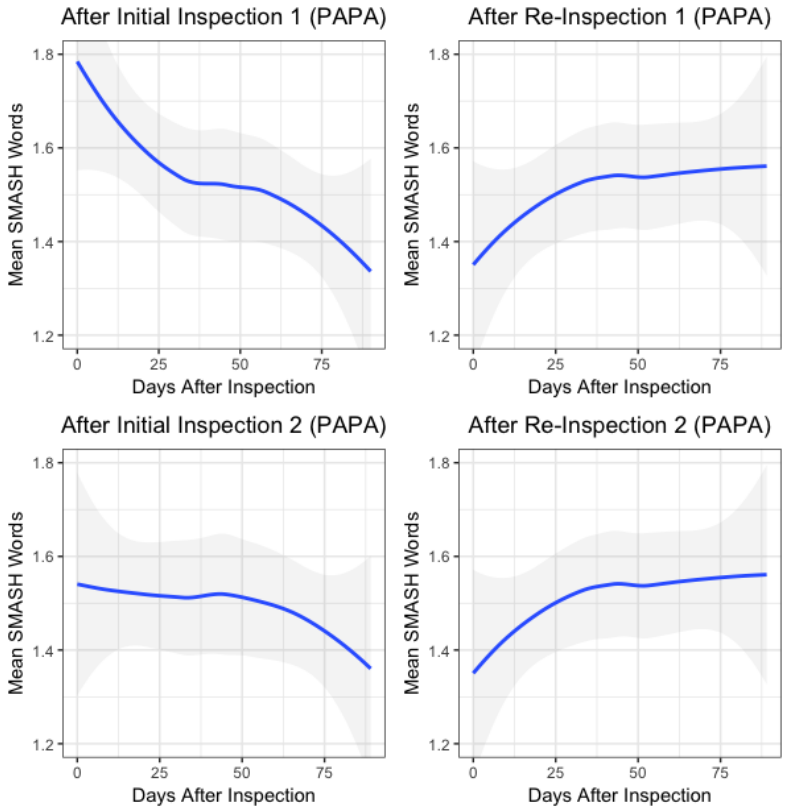


Figure A6. Comparable to Figure 3 in the main paper with 6-month interval

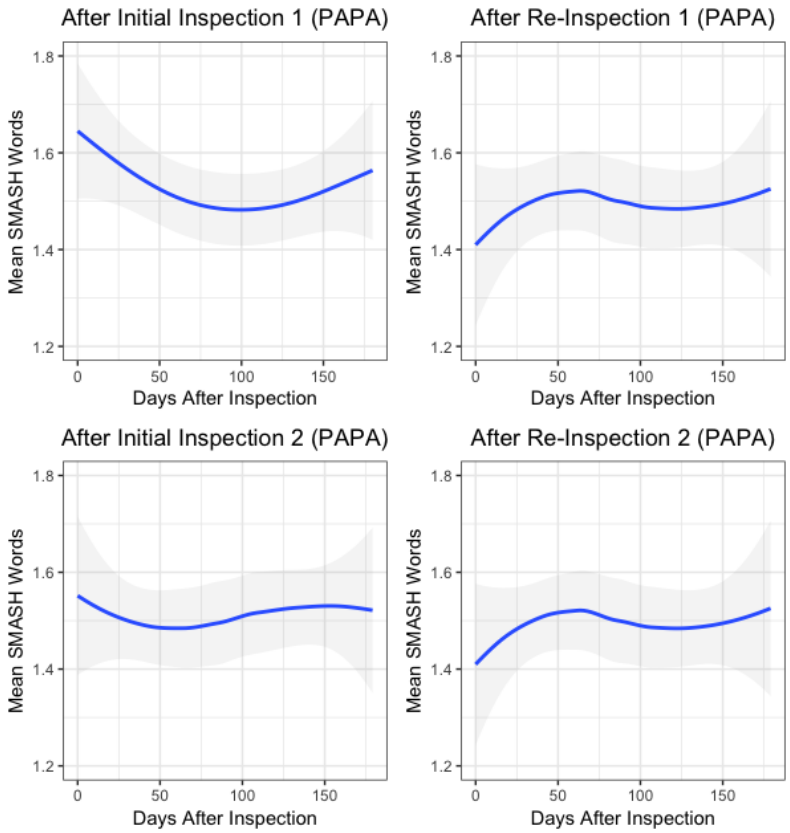
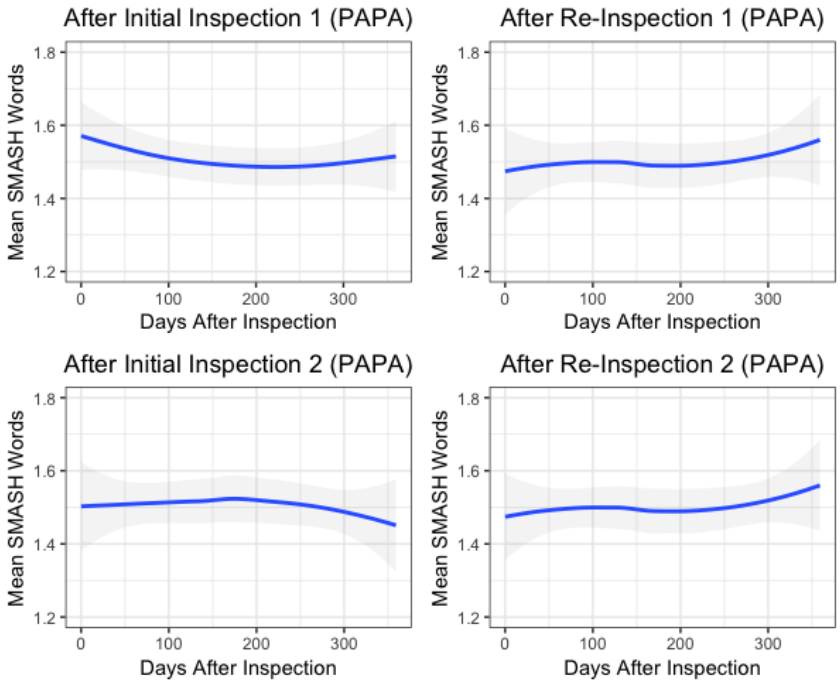


Figure A7. Comparable to Figure 3 in the main paper with 1-year interval



Appendix References

Angrist JD, Pischke JS (2009) *Mostly harmless econometrics: an empiricist's companion* (Princeton university press Princeton).

Athey S, Imbens GW (2006) Identification and inference in nonlinear difference-in-differences models. *Econometrica* 74(2):431–497.

Autor DH (2003) Outsourcing at will: The contribution of unjust dismissal doctrine to the growth of employment outsourcing. *J. Labor Econ. January*.

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B Methodol.*:289–300.

Bertrand M, Duflo E, Mullainathan S (2002) *How much should we trust differences-in-differences estimates?* (National Bureau of Economic Research).

Breiman L (2001a) Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). *Stat. Sci.* 16(3):199–231.

Breiman L (2001b) Random Forests. *Mach. Learn.* 45(1):5–32.

DrivenData (2016) Competition: Keeping it Fresh: Predict Restaurant Inspections. Retrieved https://www.drivendata.org/competitions/5/.

Hastie T, Tibshirani R, Friedman J (2009) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition* 2nd ed. (Springer-Verlag, New York).

Hoerl AE, Kennard RW (2000) Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* 42(1):80–86.

Imbens G, Wooldridge J (2007) Difference-in-differences estimation. *Natl. Bur. Econ. Res. Work. Pap.*

Imbens GW, Wooldridge JM (2009) Recent developments in the econometrics of program evaluation. *J. Econ. Lit.* 47(1):5–86.

Jin GZ, Leslie P (2009) Reputational incentives for restaurant hygiene. *Am. Econ. J. Microecon.* 1(1):237–267.

Kang JS, Kuznetsova P, Luca M, Choi Y (2013) Where Not to Eat? Improving Public Policy by Predicting Hygiene Inspections Using Online Reviews. *EMNLP*:1443–1448.

Lechner M (2011) *The estimation of causal effects by difference-in-difference methods* (Now).

Li X, Hitt LM (2008) Self-Selection and Information Role of Online Product Reviews. *Inf. Syst. Res.* 19(4):456–474.

Liaw A, Wiener M (2002) Classification and Regression by randomForest. *R News* 2(3):18–22.

Ljung GM, Box GE (1978) On a measure of lack of fit in time series models. *Biometrika* 65(2):297–303.

Muchnik L, Aral S, Taylor SJ (2013) Social Influence Bias: A Randomized Experiment. *Science* 341(6146):647–651.

New York City Department of Health and Mental Hygiene (2016a) *Food Service Establishment Inspection Scoring Parameters: A Guide to Conditions*

New York City Department of Health and Mental Hygiene (2016b) *How We Score and Grade*

New York City Department of Health and Mental Hygiene (2016c) *Self-Inspection Worksheet for Food Service Establishments*

New York City Department of Health and Mental Hygiene (2016d) *What to Expect When You're Inspected: A Guide for Food Service Operators*

New York City Department of Health and Mental Hygiene The Impact of Letter Grading in NYC.

R Core Team (2017) *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, Austria).

Schomberg JP, Haimson OL, Hayes GR, Anton-Culver H (2016) Supplementing public health inspection via social media. *PloS One* 11(3):e0152117.

Seber GAF, Lee AJ (2003) *Linear Regression Analysis* 2 edition. (Wiley, Hoboken, N.J).

Simon N, Friedman J, Hastie T, Tibshirani R (2011) Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent. *J. Stat. Softw.* 39(5):1–13.

Tibshirani R (1996) Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B Methodol.*:267–288.

Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B Stat. Methodol.* 67(2):301–320.