Studies in Knowledge and Innovation
Vol. 2 No. 11 (2026)

When Tail-Risk Estimates Leave Weekly Review Priorities Unchanged

Roger Smith · Olivia Hartman · David Miller

Published: November 26, 2026 | Article 1137

Abstract

Does giving extra weight to the worst days change which municipal request categories receive review time? We used three years of New York City 311 counts to compare rules for placing two hypothetical review blocks among four transportation request categories. The score penalizes the largest excess in any category on any day of the week, with extra weight above a calibration threshold. On 104 held-out weeks, empirical, moment-based and blended rolling rules chose the same pair every time. Their mean score was 141.36 compared with 147.57 for a fixed plan. The fixed plan scored worse in this replay, but none of the three ways of modeling recent demand changed the selected pair.

When Tail-Risk Estimates Leave Weekly Review Priorities Unchanged

Abstract

Does giving extra weight to the worst days change which municipal request categories receive review time? We used three years of New York City 311 counts to compare rules for placing two hypothetical review blocks among four transportation request categories. The score penalizes the largest excess in any category on any day of the week, with extra weight above a calibration threshold. On 104 held-out weeks, empirical, moment-based and blended rolling rules chose the same pair every time. Their mean score was 141.36 compared with 147.57 for a fixed plan. The fixed plan scored worse in this replay, but none of the three ways of modeling recent demand changed the selected pair.

Keywords sample average approximation, moment-based optimization, weekly planning, municipal service requests, decision evaluation

1. Introduction

A city office with a limited number of additional review assignments may have to decide where to place them before it sees the flow of incoming reports. The categories with the most work in a typical week need not be the ones that drive occasional extremes. A planner concerned about those extremes might evaluate the options using the empirical tail of recent weeks or introduce a distributional bound that uses additional features of the small sample. The two approaches can produce different estimates of the cost of an action while preserving the minimum, especially when the available decisions are few and several options protect the same largest potential overload.

The context for this comparison is New York City's public archive of 311 service requests. Nam and Pardo (2014) discuss how a 311 contact center integrates access to municipal services. White and Trump (2018) show that published reports can reveal differences in how residents interact with their government. Cook et al. (2024) use road reports to study heterogeneity between underlying service conditions and demand expressed by residents. In New York, Kontokosta and Hong (2021) compare reported street conditions with independent assessments and find evidence of geographically uneven reporting. The fact that requests arrive and receive a label makes them an observable input for review planning. Their counts have a narrower meaning than a census of deteriorated assets. We keep the outcome close to the record and examine a proposed unit of hypothetical administrative review.

The study asks whether augmenting sample average approximation with a moment-based tail estimate changes which two categories receive an extra review block and leads to a lower loss on later weeks. Jin et al. (2026) develop a formal approach that harmonizes sample average approximation and distributionally robust optimization. That motivates measuring the consequences of treating recent empirical weeks and partial distributional information differently. Our experiment uses a small, enumerated action set and a simple plug-in mixture in the tail term. The central unit is a joint week of four category counts observed for seven days. The response to uncertainty consists of selecting a pair of categories for the entire week. An additional aim is to distinguish gains from updating any decision using recent evidence from gains specifically attributable to the way the recent tail is estimated.

2. Source records and pilot calibration

The source is the NYC Open Data registry for 311 service requests from 2020 onward (City of New York, n.d.-a). We select the Department of Transportation agency code and four labels from the API field named complaint_type: Sidewalk Condition, Street Condition, Street Light Condition, and Traffic Signal Condition. The agency continues to use this field name in the API even though its displayed label changed to Problem (City of New York, n.d.-b). The query counts requests by the calendar date of the recorded creation timestamp and their assigned category. It uses only citywide aggregates, discarding any individual or location details. The downloaded table contains 4,384 grouped entries. Each of 1,096 calendar dates from January 1, 2023, through December 31, 2025, has a positive published count in each of the four categories. There are no missing or duplicate group keys. The work uses a retrospective snapshot, so later changes to dates, labels or the archive may alter a new extract.

The count belongs to the date on which a service request was created. It gives no indication of the agency's actual review rate, time per request, or remaining queue. Work on administrative reporting addresses additional issues that a count alone leaves undetermined. Boxer et al. (2025) estimate missing reports in one area of the NYC archive. Leo and Pate (2025) study incentives to report housing conditions. Rosłon (2026) considers the different problem of classifying infrastructure-related reports near road works using additional spatial and weather information. These studies motivate treating the series as a charge to the hypothetical planning exercise while leaving physical need, demand not reported to the city and service outcomes outside the calculation. Call-center work makes the corresponding distinction between the arrival process and the serving process explicit (Brown et al., 2005; Gans et al., 2003). Our backtest has the former counts in aggregate and does not estimate waiting times.

We partition the observations into complete Monday-through-Sunday weeks. The 52 weeks beginning January 2 through December 25, 2023, are the pilot. The 104 subsequent full weeks are the replay. The first evaluation begins January 1, 2024, and the last starts December 22, 2025, ending December 28. The remaining incomplete week buckets are outside the comparison. Each week's year is assigned by its Monday date. The pilot supplies a reference amount of daily attention for each category equal to the median recorded count rounded up. In the category order above the daily nominal capacities are 70, 173, 83 and 105. For each of the six possible pairs, adding a block of 20 to each selected category increases the total capacity from 431 to 471 record-equivalents per day. A plan stays in force for all seven days, with no carry-over of unused daily capacity across categories or dates.

The unit of a block is a hypothetical ability to review 20 records during a day. It is neither a staffing headcount nor an estimate of the service effort required by the category. The 20-unit size is a design choice and the comparison varies it later. We apply the same action sizes and nominal base capacities to every fitted rule. To select a value separating ordinary and unusually large residual work, we compute the largest positive category-day excess for each pilot week before assigning any additional blocks. The higher observed 75th percentile of those 52 maxima is 169 records. This value supplies the tail threshold for every rule, action size, history length and evaluated period. Using the base case for that calculation keeps the threshold fixed even when the two extra blocks are subsequently placed differently.

3. Empirical and moment-based rules

Each week's demand scenario is a seven-by-four matrix of recorded requests. For an action that defines four daily capacities, subtract the capacities from the scenario and take the positive part in each cell. Let Z denote the largest such positive value across that week's 28 cells. The realized score is Z plus the positive part of Z minus 169. This adds a second unit of penalty to every request above the tail threshold in the worst category-day. The score is a count-equivalent measure by construction. All categories enter with the same unit weight and the maximum permits only the worst cell to set the score. It is different from counting the week's total residual work. Stochastic programming evaluates choices over scenarios while keeping the chosen objective explicit (Birge & Louveaux, 2011; Shapiro et al., 2021). Formulations based on positive-part loss terms also appear in tail-risk analysis (Rockafellar & Uryasev, 2000). Here the threshold is fixed by the pilot, and the loss has no conditional-value-at-risk calibration.

The fixed rule minimizes the empirical mean of the stated score over all 52 pilot weeks and keeps the selected pair unchanged. The rolling sample average approximation, or SAA, rule computes the same mean for each of the six pairs using the previous eight complete weeks. Kleywegt et al. (2002) describe the sample average approximation approach for discrete decisions, and Mak et al. (1999) distinguish selected solutions from their expected quality on new draws. The floating history supplies eight joint scenarios at every origin and preserves each scenario's day and category composition. It also exposes the empirical tail term to the inclusion or exclusion of just one unusually large week. The analysis uses the observed chronology to compare decisions before and after such changes.

For the rolling moment rule let m and s denote the mean and standard deviation of the eight Z values for one pair, using divisor eight for the variance. The base mean term remains m. In the tail term we substitute the bound $0.5 \times [(m - 169) + \sqrt{((m - 169)^2 + s^2)}]$. This follows by writing the positive part of a centered loss as half of its signed value plus half of its absolute value, and bounding the expectation of the absolute value by its root second moment. It is a sharp upper bound for expected exceedance over real-valued distributions with those moments, with equality possible at a suitable two-point distribution. Because the true weekly maximum excess is nonnegative, this bound may be loose when a maximizing distribution puts mass below zero. The calculation is consistent with distribution-free positive-part loss bounds (Gallego & Moon, 1993) and the use of moment-based optimization bounds studied by Bertsimas and Popescu (2005). Crucially, moments are estimated separately for the action-specific Z values after taking each week's maximum. No assumption of independent days or categories is needed to construct those values.

The third rolling rule averages the empirical tail expectation and the moment envelope and then adds the same mean term. This blend applies a fixed one-half weight without tuning on the held-out results. Delage and Ye (2010) model uncertainty in fitted moments themselves, and Wiesemann et al. (2014) present broader constructions for ambiguity sets. Statistical calibration can also be applied to robust constraints or to distances from an empirical distribution (Bertsimas et al., 2018; Mohajerin Esfahani & Kuhn, 2018). Ben-Tal et al. (2013) consider uncertainty in probabilities for a given set of scenarios. Our calculation uses fixed estimated moments and a closed-form scenario-based adjustment, which makes it possible to see exactly when the extra penalty changes the decision. It entails no guarantee on future scores. Each optimizer enumerates six candidate pairs and resolves equal fitted costs in the same category order.

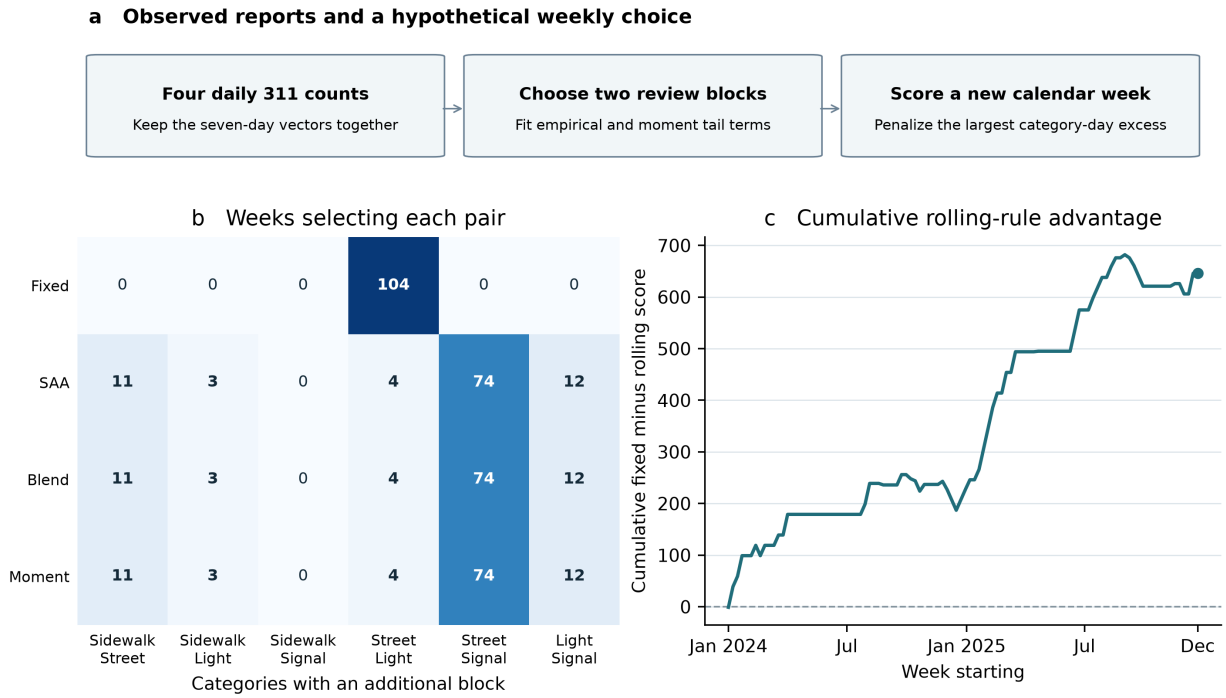


Figure 1. *Panel a shows the weekly comparison. Panel b gives observed counts of the chosen pairs across held-out origins. Panel c plots the cumulative score difference between the fixed comparator and any one of the three rolling rules at the main specification, where all three rolling rules have identical scores.*

Chronological replay uses the preceding completed weeks only. Tashman (2000) describes rolling origins for out-of-sample tests, and Cerqueira et al. (2020) compare time-series performance-estimation procedures under different data conditions. We report the mean realized score, the mean worst category-day excess, threshold crossings, and an oracle that picks the best of the same six choices after seeing the week. The latter only contextualizes the size of gains available within the modeled action set. Pairing each policy's loss on the same realized weeks follows the loss-differential approach in Diebold and Mariano (1995). The descriptive interval compares average scores rather than using the conditional testing framework of Giacomini and White (2006). We form 95% percentile intervals using 10,000 draws of four-week circular blocks, and repeat the computation with eight-week blocks. Künsch (1989) and Politis and Romano (1994) provide different approaches to preserving serial dependence by resampling time blocks. The present intervals make the block choice visible and describe a short archive with possibly changing request patterns.

4. Measured replay

The three rolling methods selected the same pair on all 104 held-out weeks at the main specification. Each put both blocks into Street Condition and Traffic Signal Condition on 74 weeks. Street Condition was paired with Sidewalk Condition on 11 weeks and with Street Light Condition on four. The remaining 15 weeks used Street Light Condition with Traffic Signal Condition or Sidewalk Condition. The fixed comparator selected Street Condition and Street Light Condition for all weeks, reflecting the lowest empirical score on the pilot.

Table 1. *The same 104 weeks under the 20-record block and eight-week lookback. The fixed rule is chosen during calibration. Each rolling rule chooses a pair using information from previously completed weeks.*

Rule	Mean realized score	Mean largest excess (records)	Weeks above 169 records	90th percentile of largest excess (records)
Calibration-fixed	147.57	126.73	22	234.4
Rolling SAA	141.36	122.94	19	227.6
Rolling blend	141.36	122.94	19	227.6
Rolling moment	141.36	122.94	19	227.6

The mean gap in score between the fixed and rolling rules is 6.21. The paired four-week resampling interval for fixed minus rolling is 2.45 to 10.28. With eight-week blocks it is 2.08 to 10.93. The oracle averages 137.48, which is 3.88 lower than the rolling methods and 10.09 lower than the fixed rule. The maximum realized worst excess is 492 for every main rule. The update helps on average in this archive while leaving the largest realized excess intact. In 2024 the fixed and rolling mean scores are 123.00 and 118.72, with eight threshold crossings for both. In 2025 the corresponding scores are 173.10 and 164.88, with 14 and 11 crossings. The years differ in overall score scale, and the improvement in mean score has the same sign in each.

Despite identical actions, the moment calculation changes the full ordering of the six pairs on 11 weeks. Across origins the median within-week spread in the added moment penalty from one pair to another is 1.97 score units. Even where the ranking farther from the minimum changes, the empirical and moment rules retain the same sets of minimizing choices in all main weeks. There is also substantial coarseness in the realized problem. With hindsight, 75 of 104 weeks have three loss-minimizing plans, and 29 have a unique one. Often the action must place a block on the category that generates the largest excess, and the second assigned block can go to more than one place without changing the maximum. Adding a distinct tail estimate can change the fitted cost of suboptimal alternatives while leaving the actionable difference between the leading alternatives at zero.

Table 2. *Sensitivity in mean realized score. The same archive, date boundaries, base capacities, weekly objective and pilot tail threshold are used for every row. Block size changes the objective under all policies applied within that row.*

Additional capacity per selected category	Rolling weeks	Fixed	SAA	Blend	Moment
10 records	8	153.78	151.08	151.08	150.81
20 records	4	147.57	142.26	141.07	141.07
20 records	16	147.57	141.51	141.13	141.13
40 records	8	137.85	124.93	124.81	124.81

The most visible departure from the main tie is at the four-week history. The blend and moment rules each average 1.19 score units less than the rolling SAA rule. At 16 weeks the improvement in this sample is 0.38, and at a 40-record block it is 0.13. At a 10-record block the moment rule averages slightly less than SAA even though the blend has the same score as SAA at displayed precision. Different actions can also lead to exactly the same realized score on some weeks. Across the chosen sensitivity ranges, changing the number of additional records affects the score much more than altering the tail calculation. Those differences are properties of the hypothetical capacity and objective, and do not establish the benefits or cost-effectiveness of larger staffing assignments.

5. Discussion

The result is best understood at the level of the decision. When a week is judged by its worst category-day combination, a candidate pair earns a better score only if one of its two blocks covers a binding or near-binding category. A change to the estimated tail of the resulting weekly loss may leave this part of the comparison intact. This limits the opportunity for a more elaborate empirical-distribution treatment to improve choices in a small menu. It is a different observation from finding that both tail estimators are equally accurate for all actions. The ranking audit shows that the choice to use moments matters for the estimated alternatives, yet the difference is below the level at which a new decision is selected in the main window.

The comparison has direct limits. The group of requests may contain records for the same physical incident, and historical labels can shift. No observations on review times, backlog carried between days, or how extra capacity could alter reporting behavior are used. The use of 20 review equivalents per day and the doubling of the marginal score above 169 are analytical choices. A different loss such as the sum of all category-day excesses would give each additional reviewed record a different value and could favor allocating attention to jointly elevated categories even when one category dominates the peak. This study gives one situation in which separating empirical tail estimation from action evaluation makes a substantive difference. Rolling selection improves on the fixed historical pair in mean replay score, and the common action sequence shows that this improvement is available without making a tail-estimator choice within the tested main design.

Data and code availability

The City of New York's public table supplies the underlying counts. The accompanying grouped query fetches counts only. The analysis exports the archived aggregate input hash, evaluation scores by week and rule, and results for the sensitivity and choice audit. The study figure is generated from those results.

References

Ben-Tal, A., den Hertog, D., De Waegenaere, A., Melenberg, B., & Rennen, G. (2013). Robust solutions of optimization problems affected by uncertain probabilities. *Management Science, 59*(2), 341–357. https://doi.org/10.1287/mnsc.1120.1641

Bertsimas, D., Gupta, V., & Kallus, N. (2018). Data-driven robust optimization. *Mathematical Programming, 167*(2), 235–292. https://doi.org/10.1007/s10107-017-1125-8

Bertsimas, D., & Popescu, I. (2005). Optimal inequalities in probability theory: A convex optimization approach. *SIAM Journal on Optimization, 15*(3), 780–804. https://doi.org/10.1137/S1052623401399903

Birge, J. R., & Louveaux, F. (2011). *Introduction to stochastic programming* (2nd ed.). Springer. https://doi.org/10.1007/978-1-4614-0237-4

Boxer, K. S., Hong, B., Kontokosta, C. E., & Neill, D. B. (2025). Estimating reporting bias in 311 complaint data. *The Annals of Applied Statistics, 19*(2), 1691–1713. https://doi.org/10.1214/24-AOAS2003

Brown, L., Gans, N., Mandelbaum, A., Sakov, A., Shen, H., Zeltyn, S., & Zhao, L. (2005). Statistical analysis of a telephone call center: A queueing-science perspective. *Journal of the American Statistical Association, 100*(469), 36–50. https://doi.org/10.1198/016214504000001808

Cerqueira, V., Torgo, L., & Mozetič, I. (2020). Evaluating time series forecasting models: An empirical study on performance estimation methods. *Machine Learning, 109*(11), 1997–2028. https://doi.org/10.1007/s10994-020-05910-7

City of New York. (n.d.-a). *311 service requests from 2020 to present* [Dataset]. NYC Open Data. https://data.cityofnewyork.us/Social-Services/311-Service-Requests-from-2010-to-Present/erm2-nwe9/data

City of New York. (n.d.-b). *311 service requests updates*. NYC Open Data. https://www.nyc.gov/opendata/news/all-news/311-Service-Requests-Updates

Jin, Z., Cheng, J., Long, D. Z., & Pan, K. (2026). Harmonizing sample average approximation and distributionally robust optimization. *Omega, 144*, 103584. https://doi.org/10.1016/j.omega.2026.103584

Cook, S. J., Zuhlke, S., & Saywitz, R. (2024). Potholes, 311 reports, and a theory of heterogeneous resident demand for city services. *Policy Studies Journal, 52*(3), 647–669. https://doi.org/10.1111/psj.12540

Delage, E., & Ye, Y. (2010). Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations Research, 58*(3), 595–612. https://doi.org/10.1287/opre.1090.0741

Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics, 13*(3), 253–263. https://doi.org/10.1080/07350015.1995.10524599

Gallego, G., & Moon, I. (1993). The distribution free newsboy problem: Review and extensions. *Journal of the Operational Research Society, 44*(8), 825–834. https://doi.org/10.1057/jors.1993.141

Gans, N., Koole, G., & Mandelbaum, A. (2003). Telephone call centers: Tutorial, review, and research prospects. *Manufacturing & Service Operations Management, 5*(2), 79–141. https://doi.org/10.1287/msom.5.2.79.16071

Giacomini, R., & White, H. (2006). Tests of conditional predictive ability. *Econometrica, 74*(6), 1545–1578. https://doi.org/10.1111/j.1468-0262.2006.00718.x

Kleywegt, A. J., Shapiro, A., & Homem-de-Mello, T. (2002). The sample average approximation method for stochastic discrete optimization. *SIAM Journal on Optimization, 12*(2), 479–502. https://doi.org/10.1137/S1052623499363220

Kontokosta, C. E., & Hong, B. (2021). Bias in smart city governance: How socio-spatial disparities in 311 complaint behavior impact the fairness of data-driven decisions. *Sustainable Cities and Society, 64*, 102503. https://doi.org/10.1016/j.scs.2020.102503

Künsch, H. R. (1989). The jackknife and the bootstrap for general stationary observations. *The Annals of Statistics, 17*(3), 1217–1241. https://doi.org/10.1214/aos/1176347265

Leo, G., & Pate, J. (2025). Complainer's dilemma. *Journal of Public Economic Theory, 27*(1), e70011. https://doi.org/10.1111/jpet.70011

Mak, W.-K., Morton, D. P., & Wood, R. K. (1999). Monte Carlo bounding techniques for determining solution quality in stochastic programs. *Operations Research Letters, 24*(1–2), 47–56. https://doi.org/10.1016/S0167-6377(98)00054-6

Mohajerin Esfahani, P., & Kuhn, D. (2018). Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming, 171*(1–2), 115–166. https://doi.org/10.1007/s10107-017-1172-1

Nam, T., & Pardo, T. A. (2014). Understanding municipal service integration: An exploratory study of 311 contact centers. *Journal of Urban Technology, 21*(1), 57–78. https://doi.org/10.1080/10630732.2014.887933

Politis, D. N., & Romano, J. P. (1994). The stationary bootstrap. *Journal of the American Statistical Association, 89*(428), 1303–1313. https://doi.org/10.1080/01621459.1994.10476870

Rockafellar, R. T., & Uryasev, S. (2000). Optimization of conditional value-at-risk. *The Journal of Risk, 2*(3), 21–41. https://doi.org/10.21314/JOR.2000.038

Rosłon, J. (2026). Risk prioritization of nearby infrastructure-related 311 requests around urban streetworks using public data: A weather-informed calibrated random forest case study in New York City. *Infrastructures, 11*(8), 261. https://doi.org/10.3390/infrastructures11080261

Shapiro, A., Dentcheva, D., & Ruszczyński, A. (2021). *Lectures on stochastic programming: Modeling and theory* (3rd ed.). Society for Industrial and Applied Mathematics. https://doi.org/10.1137/1.9781611976595

Tashman, L. J. (2000). Out-of-sample tests of forecasting accuracy: An analysis and review. *International Journal of Forecasting, 16*(4), 437–450. https://doi.org/10.1016/S0169-2070(00)00065-0

White, A., & Trump, K.-S. (2018). The promises and pitfalls of 311 data. *Urban Affairs Review, 54*(4), 794–823. https://doi.org/10.1177/1078087416673202

Wiesemann, W., Kuhn, D., & Sim, M. (2014). Distributionally robust convex optimization. *Operations Research, 62*(6), 1358–1376. https://doi.org/10.1287/opre.2014.1314