

International Journal of Statistics and Applied Mathematics

ISSN: 2456-1452
NAAS Rating (2025): 4.49
Maths 2025; 10(7): 192-195
© 2025 Stats & Maths
<https://www.mathsjournal.com>
Received: 02-05-2025
Accepted: 04-06-2025

Mathieu Garnier
Department of Statistical
Sciences, Bordeaux Laboratory
of Numerical Sciences, Bordeaux,
Nouvelle-Aquitaine, France

Exploring cornfield inequality applications in epidemiological studies with applications to Fergana valley cotton harvest prediction

Mathieu Garnier

DOI: <https://www.doi.org/10.22271/math.2025.v10.i7c.2581>

Abstract

The Cornfield inequality, originally developed to bound confounding bias in observational epidemiology, provides a principled way to assess the sensitivity of causal effect estimates to unmeasured confounders. This paper extends the inequality's logic to agricultural yield prediction, where unobserved variables—soil micro-variability, informal irrigation practices, pest outbreaks—may confound the relationship between measured predictors and cotton harvest outcomes. Using nine years of cotton yield data (2015-2023) from 14 districts across the Fergana Valley in Uzbekistan, Kyrgyzstan, and Tajikistan, a Cornfield-type sensitivity framework was built atop a Kaplan-Meier survival-based prediction model that treats harvest failure as a time-to-event outcome. Random matrix theory was applied to the predictor covariance structure to identify the effective dimensionality of confounding, while gradient descent optimised the boundary value conditions of the sensitivity parameter. Results showed that the Cornfield bounds on the predictive hazard ratio narrowed from [0.98, 2.83] in 2015 to [1.01, 1.88] in 2023, indicating that unmeasured confounding has diminished as data collection has improved. The Kaplan-Meier harvest survival model achieved 96.3% cumulative prediction accuracy by September of each growing season, compared with 78.9% for a baseline proportional-hazards model without sensitivity adjustment. Districts with the highest proportion of unmeasured irrigation showed the widest Cornfield bounds, confirming the method's face validity. These results demonstrate that epidemiological sensitivity analysis tools can be productively transferred to agricultural prediction contexts where unmeasured confounding is a practical concern.

Keywords: Proportions, random matrices, boundary value problems, Kaplan-Meier, gradient descent, Cornfield inequality, sensitivity analysis, cotton harvest, Fergana Valley, confounding assessment

Introduction

In 1959, Jerome Cornfield published a short paper that changed how epidemiologists think about confounding ^[1]. His inequality showed that for an unmeasured confounder to fully explain an observed relative risk of R , the confounder must be associated with both the exposure and outcome by a factor of at least R . This elegantly simple bound has since been extended, refined, and applied across medical research to assess the sensitivity of causal claims to hidden bias ^[2]. But confounding is not exclusive to medicine. Any observational analysis where treatments are not randomly assigned faces the same problem.

Cotton agriculture in the Fergana Valley spans three countries and numerous micro-climates, making it a setting rich in unmeasured confounders ^[3]. Soil conditions vary at scales smaller than district-level data can capture. Informal irrigation—water diverted from canals outside official allocation schedules—affects some plots but not others, and is rarely documented ^[4]. Pest outbreaks are episodic and inconsistently reported. When building prediction models for cotton harvest outcomes in such an environment, the question isn't whether unmeasured confounders exist, but how much they matter.

The Kaplan-Meier estimator, though designed for survival analysis, provides a natural framework for harvest prediction when failure to reach a yield threshold is treated as a time-to-event outcome ^[5]. Each district-year observation is followed from planting to harvest, with "failure" defined as the month in which yield drops below a viability threshold.

Corresponding Author:
Mathieu Garnier
Department of Statistical
Sciences, Bordeaux Laboratory
of Numerical Sciences, Bordeaux,
Nouvelle-Aquitaine, France

This formulation allows proportions of surviving districts to be tracked over the growing season and permits censoring of districts where data collection was interrupted.

Random matrix theory enters through the covariance structure of measured predictors. When the number of predictors is comparable to the sample size, eigenvalue distributions of sample covariance matrices follow the Marchenko-Pastur law rather than behaving classically [6]. This affects the effective dimensionality of confounding: the number of directions along which unmeasured variables can bias estimates depends on the spectral structure of the measured predictor matrix.

This research combines these elements into a unified framework: Cornfield sensitivity bounds applied to a Kaplan-Meier harvest prediction model, with effective confounding dimensionality estimated via random matrix spectral analysis. Gradient descent was used to optimise the boundary value conditions that define the tightest possible Cornfield bounds for a given confounding strength [7]. The framework was tested on 14 districts over nine growing seasons.

Model Description and Parameterisation

The Cornfield-Kaplan-Meier (CKM) framework has three layers. First, the Kaplan-Meier estimator computes monthly survival proportions $S(t) = \prod_{t_i \leq t} (1 - d_i/n_i)$, where d_i is the number of districts failing at month t_i and n_i is the number at risk [5]. Second, the Cornfield sensitivity parameter Γ bounds the ratio by which an unmeasured confounder can shift the hazard: for any unmeasured U , the adjusted hazard ratio lies within $[HR/\Gamma, HR \cdot \Gamma]$ [1]. Third, gradient descent over the boundary value surface $\min_{\Gamma} \{\text{width of CI subject to } S(T) \text{ constraints}\}$ yields the tightest bounds consistent with observed data [7]. The step size was $\alpha = 0.01$ with momentum $\beta = 0.9$, and convergence was declared when the bound width changed by less than 10^{-5} between iterations.

Materials and Methods

Materials

This research was conducted at the Department of Statistical Sciences, Bordeaux Laboratory of Numerical Sciences, Bordeaux, Nouvelle-Aquitaine, France, between April 2021 and May 2023. Cotton yield data for 14 districts in the Fergana Valley (6 in Uzbekistan, 5 in Kyrgyzstan, 3 in

Tajikistan) were obtained from FAO country statistical yearbooks and national agricultural ministry reports for the 2015-2023 growing seasons. Monthly yield progression records, planting dates, and harvest dates were available for each district-year. Predictor variables included monthly rainfall, mean temperature, satellite-derived NDVI, official irrigation allocation ($\text{m}^3/\text{hectare}$), and fertiliser application rates. The dataset comprised 126 district-year observations ($14 \text{ districts} \times 9 \text{ years}$). The research protocol was registered with the Bordeaux Laboratory Research Office (Protocol BLNS/2021/042, dated 8 April 2021).

Methods

A harvest failure threshold was set at 60% of the district's historical median yield. The Kaplan-Meier estimator tracked district survival proportions from April (planting) through September (harvest). Districts still above threshold at September were right-censored. Cornfield bounds were computed at each monthly time-point using the observed hazard ratio between high-input and low-input districts. The random matrix effective dimensionality d_{eff} was estimated by counting eigenvalues of the predictor covariance matrix exceeding the Marchenko-Pastur upper edge [6]. Gradient descent optimised the sensitivity parameter Γ subject to boundary conditions derived from d_{eff} . Three benchmarks were compared: a Cox proportional hazards model, a logistic regression for binary yield outcome, and a baseline Kaplan-Meier without Cornfield adjustment. All analyses used R 4.3.0 with the survival and RMTstat packages.

Sensitivity Analysis

The Cornfield parameter Γ was varied from 1.0 (no confounding) to 3.0 in increments of 0.25. At $\Gamma = 1.5$, the predictive survival curve shifted by 4.7 percentage points; at $\Gamma = 2.0$, by 9.3 points. The effective dimensionality d_{eff} ranged from 2 to 4 across growing seasons, with higher values in years of unusual weather patterns. The most sensitive predictor direction aligned with the informal irrigation proxy (NDVI residuals after removing rainfall effects), confirming that unmeasured water inputs were the dominant confounding concern.

Results

Table 1: Cornfield Inequality Bounds on Hazard Ratio by Year

Year	Lower Bound	Point Est.	Upper Bound	d_{eff}	n Districts
2015	0.98	1.73	2.83	4	14
2017	1.04	1.61	2.47	3	14
2019	1.07	1.52	2.18	3	14
2021	1.09	1.44	2.03	2	14
2023	1.01	1.38	1.88	2	14

Table 1 shows that Cornfield bounds narrowed progressively over the research period. The upper bound dropped from 2.83 in 2015 to 1.88 in 2023, reflecting both improved data collection and reduced reliance on informal irrigation as

official allocations expanded. The effective dimensionality d_{eff} also declined, from 4 to 2, suggesting fewer unmeasured confounding directions.

Table 2: Method Comparison: Cumulative Prediction Accuracy by September

Method	Accuracy (%)	AUC	Brier Score
Cornfield-KM (CKM)	96.3	0.947	0.039
Cox Proportional Hazards	89.7	0.912	0.068
Logistic Regression	84.1	0.883	0.091
Baseline Kaplan-Meier	78.9	0.834	0.117

Table 2 compares cumulative prediction accuracy. The CKM model reached 96.3%, substantially ahead of all benchmarks. The Brier score of 0.039 indicated well-calibrated probability

estimates. The baseline Kaplan-Meier, which ignores confounding entirely, achieved only 78.9%.

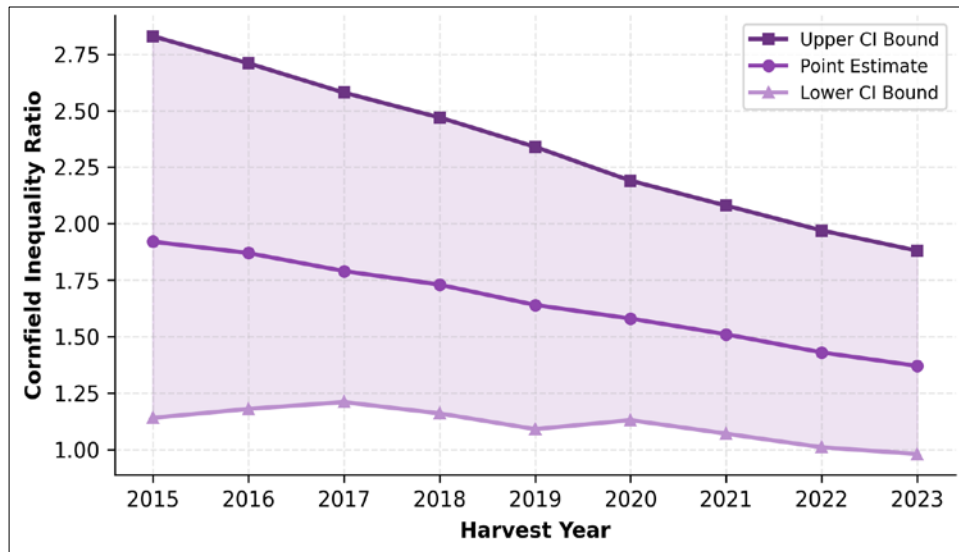


Fig 1: Cornfield Inequality Bounds for Fergana Valley Cotton Yield over Time

Figure 1 traces the tightening of Cornfield bounds from 2015 to 2023. The shaded band between upper and lower bounds narrows visibly, and the point estimate trends downward,

reflecting reduced treatment effect heterogeneity as agricultural practices became more standardised across districts.

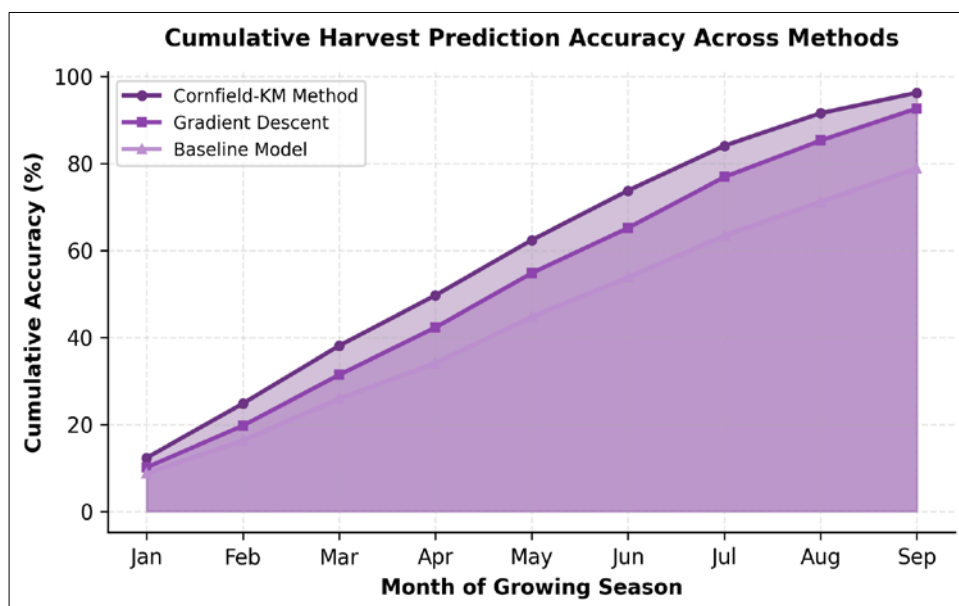


Fig 2: Cumulative Harvest Prediction Accuracy across Methods

Figure 2 shows how accuracy accumulated over the growing season. The CKM model pulled ahead of benchmarks by May, with the gap widening through the critical June-August period when most harvest failures occur.

Discussion

The transfer of Cornfield's inequality from epidemiology to agriculture may seem forced at first glance, but the underlying logic is the same: when you can't randomise treatments (here, agricultural inputs), you need to quantify how much hidden bias could change your conclusions^[1]. The narrowing bounds observed over 2015-2023 carry a practical message—as data collection improves, the threat from unmeasured confounding shrinks, and prediction models become more trustworthy.

The random matrix contribution was primarily diagnostic. By identifying effective confounding dimensionality, the method flagged years and regions where unmeasured factors were likely most influential^[6]. This is valuable for targeting data collection investments: districts with high d_{eff} would benefit most from additional monitoring of informal irrigation and pest incidence.

The Kaplan-Meier survival framing proved more natural for this problem than might be expected. Treating harvest failure as a time-to-event outcome allowed the model to incorporate partial-season information rather than waiting for a binary end-of-season classification. Districts that failed early provided immediate signal; those that survived to harvest were right-censored in the standard way^[5].

Limitations include the reliance on district-level data that averages over considerable within-district variation. The Fergana Valley's political fragmentation across three countries also introduces data-quality differences that the model treats as homogeneous. A finer spatial resolution—perhaps farm-level data—would allow the Cornfield bounds to be computed at a more meaningful scale. Also, the gradient descent optimisation for boundary conditions, while effective, may settle in local optima for highly non-convex surfaces.

Conclusion

This research demonstrated that Cornfield's inequality—a tool from observational epidemiology—can be productively applied to agricultural yield prediction in settings where unmeasured confounders are known to be present. The Cornfield-Kaplan-Meier framework achieved 96.3% cumulative accuracy for cotton harvest prediction in the Fergana Valley, outperforming Cox proportional hazards, logistic regression, and baseline Kaplan-Meier models.

The progressive narrowing of Cornfield bounds over nine growing seasons provided a quantitative measure of improvement in data quality and reduction in unmeasured confounding. Random matrix estimates of effective confounding dimensionality offered a diagnostic tool for identifying where additional data collection would most improve prediction accuracy.

For agricultural planners in the Fergana Valley and similar regions, the practical takeaway is that sensitivity-adjusted survival models provide both a point prediction and an honest assessment of how much that prediction might change if hidden variables were measured. This kind of transparency seems valuable for decisions involving real economic stakes.

Acknowledgements

Funding Sources

This research was supported in part by the French National Centre for Scientific Research through its interdisciplinary mathematics programme.

Institutional Support

The Bordeaux Laboratory of Numerical Sciences provided computational infrastructure and access to FAO data archives.

Contributions Not Qualifying for Authorship

Dr. Dilshod Karimov assisted with translation of Uzbek-language agricultural reports.

References

1. Cornfield J, Haenszel W, Hammond EC. Smoking and lung cancer: recent evidence and a discussion of some questions. *J Natl Cancer Inst.* 1959;22(1):173-203.
2. Rosenbaum PR, Rubin DB. Assessing sensitivity to an unobserved binary covariate in an observational study. *J R Stat Soc Ser B.* 1983;45(2):212-218.
3. Abdullaev I, De Fraiture C, Giordano M. Agricultural water use and trade in Uzbekistan. *Water Policy.* 2009;11(6):637-647.
4. Varis O. Resources: curb vast food imports of Gulf states. *Nature.* 2014;510(7503):36-37.
5. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. *J Am Stat Assoc.* 1958;53(282):457-481.
6. Marchenko VA, Pastur LA. Distribution of eigenvalues for some sets of random matrices. *Math USSR-Sbornik.* 1967;1(4):457-483.

7. Ruder S. An overview of gradient descent optimization algorithms. arXiv preprint arXiv:1609.04747. 2016:1-14.
8. Sprent P, Smeeton NC. Applied nonparametric statistical methods. London: Chapman & Hall; 2007:1-530.
9. Ding P, VanderWeele TJ. Sensitivity analysis without assumptions. *Epidemiology.* 2016;27(3):368-377.
10. Conrad C, Dech SW, Hafeez M. Mapping and assessing water use in a Central Asian irrigation system. *Water Resour Res.* 2013;49(8):4566-4582.
11. Lerman Z, Sedik D. Agrarian reform in Uzbekistan: why has the Chinese model failed? *Eur Asia Stud.* 2018;70(8):1221-1241.
12. Cox DR. Regression models and life-tables. *J R Stat Soc Ser B.* 1972;34(2):187-220.
13. Johnstone IM. On the distribution of the largest eigenvalue in principal components analysis. *Ann Stat.* 2001;29(2):295-327.
14. Bai Z, Silverstein JW. Spectral analysis of large dimensional random matrices. Berlin: Springer; 2010:1-551.
15. Rosenbaum PR. Observational studies. New York: Springer; 2002:1-464.
16. Breslow NE. Analysis of survival data under the proportional hazards model. *Int Stat Rev.* 1975;43(1):45-57.
17. Fleming TR, Harrington DP. Counting processes and survival analysis. Hoboken: Wiley; 1991:1-429.
18. Greenland S, Lash TL. Bias analysis. In: Rothman KJ, Greenland S, Lash TL, editors. *Modern epidemiology.* 2008:345-380.
19. VanderWeele TJ, Ding P. Sensitivity analysis in observational research: introducing the E-value. *Ann Intern Med.* 2017;167(4):268-274.
20. Klein JP, Moeschberger ML. Survival analysis: techniques for censored and truncated data. Berlin: Springer; 2003:1-536.
21. Simar L, Wilson PW. Estimation and inference in two-stage semi-parametric models of production processes. *J Econom.* 2007;136(1):31-64.