# Nonparametric and Unsupervised: NU-Learning from Big Data

Bob Obenchain, Risk-Benefit Statistics LLC, http://localcontrolstatistics.org

## Background

Nonparametric methods are commonly described as "distribution-free." Unsupervised learning tends to be "model-agnostic." *NU Learning* approaches, by being clearly antithetical to traditional parametric-supervised model-fitting, offer unique opportunities to provide data-based insights that are both unbiased and truly objective.

## NU-Learning Prototype

I have advocated a form of NU-learning, called "Local Control," for more than fifteen years [1-7]. The initial objective of "LC" Strategy is to obtain a *distribution* of unbiased "Local" Average Treatment Effect estimates (or "Local" Outcome-Exposure Associations) from *meaningful subgroups* [1] of experimental units (patients, etc.) Subgroups are formed by clustering units on their most relevant X-confounder characteristics. In this short article, I hope to stimulate interest in development of *NU* algorithms that are more efficient for truly Big Data. Specifically, I wish to pass a small "torch" forward to software developers who will, I hope, ultimately provide tools capable of cross-sectional analyses of many more than 100,000 patients. Readers interested in exploring a "subset" of their Big Data (say, at most 50K patients) can use my current R-package [5] to display key data-analytic *visualizations.*

## Unsupervised Patient Matching vs Supervised Propensity Estimation

A string of key-concept papers [8-15] provides a strong foundation for both "approximate" patient *matching* and use of the *observed propensities* (treatment-choice fractions) within the resulting patient *clusters* to reduce bias in *comparative-effectiveness research.* Again, it's *not* necessary to risk fitting any possibly "wrong" global model (e.g. a logistic regression with interaction terms) simply to provide mere propensity "estimates." After all, algorithms for *clustering* and *matching* are the most widely used *NU* methods; better and better "K-Means" (or K-Medians) algorithms should emerge over time.

The current R-implementation of LocalControl Strategy [5] uses *hierarchical* clustering, even though such methods cannot "scale up" well to truly Big datasets. But hierarchical methods are usually good for "almost" big datasets where the number of patients (experimental units) is no more than roughly N = 50,000. Once a clustering tree (dendrogram) has been computed, it becomes efficient to monitor "bias-variance trade-offs" in "local" treatment effect-size estimation as the number of Clusters requested, K, is increased from 1 to 50, to 100, ...to at most, say, N/12.

## Trade-Offs in Choice of Number of Clusters

When the smallest cluster formed contains at least 11 patients, many within-cluster statistics are (or should be) widely considered "publishable." For example, current CMS standards [16] for preserving patient privacy allow within-subgroup treatment effect-size estimates from "more than 10 patients" to be published in journals. If such detailed cluster-level information were made widely available for electronic download by publishers of research, all health-care researchers could evaluate this "supplemental information" and help in development of consensus views …especially if maximum cluster sizes are also restricted to, say, at most 20 or 30 patients. Authors and sponsors of sound research should be recognized for their data sharing efforts. In return, they retain their rights of access to and responsibility for preserving privacy of all patient-level information.

On the other hand, as illustrated below with Figures from a recent case study [4], using many *fewer* and much *larger* clusters (than 11 experimental units)

often "optimizes" clear Variance-Bias trade-offs in "local" effect-size estimation. But the extreme choice of using only K=1 "cluster" containing all N experimental units is *never* optimal; it provides only "one-size-fits-all" answers with minimum apparent variance but maximum true bias from confounded real-word data [8].

## Visualizing Local Effect-Size Distributions

Estimates of "K" local treatment effect-sizes, *weighted* proportional to cluster size, can be displayed in a simple histogram, like Figure 1.

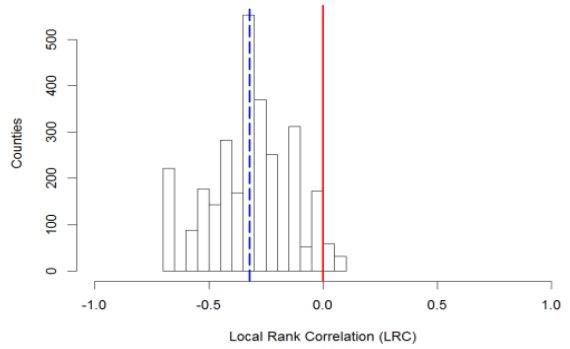


**Figure 1.** This is a histogram of local Spearman rank correlations between lung cancer mortality rates (y-outcome) and indoor radon exposure level within K = 50 clusters of 2,881 US counties [4]. The 3 X-confounders used to form clusters are percentages of county residents who (i) are over 65, (ii) currently smoke, and/or (iii) are obese. Since local associations are mostly negative here, note that local cancer mortality rates generally tend to decrease as radon exposure levels increase.

Next, displays like Figure 2 make it easy to literally "see" how initially stable or ultimately unstable the local "distribution" can become as K increases. Note that "Variance-Bias trade-offs" are illustrated using a sequence of box-and-whisker diagrams, each depicting a full "local effect-size distribution" as K is systematically increased. As outlined in the caption of Figure 2, an analyst literally "sees" that using only a few clusters (each much larger than 11 experimental units) optimizes this trade-off.

## Permutation Test: Are X-confounders Ignorable?

A local treatment effect-size distribution is truly "meaningful" [1] only if its distribution is clearly different from the purely random distribution generated by random assignment of N units to K subgroups of same sizes ($N_1$, $N_2$, …, $N_K$) as the K observed clusters. A two-sample Kolmogorov-Smirnov "D-statistic" can then be used in a permutation test, but its "p-value" must be simulated because there are *many* within-cluster ties; see Figure 3 (next page). In any case, it is rather easy to determine whether this p-value is or isn't less than 0.01. My recent presentation at MBSW [6] gives five case-study examples.

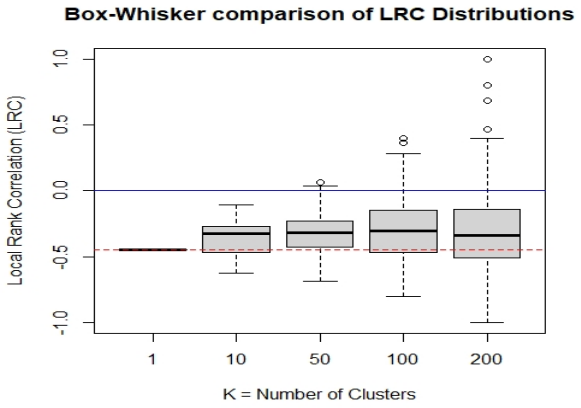


**Figure 2.** The most obvious effect of increasing K is that successive "local" effect-size distributions become more and more "spread out" vertically. Corresponding changes in (negative) LRC medians are not monotone; they start increasing towards zero for K = 10 & 50, reach their maximum at K = 100, and then start decreasing at K = 200. Ultimately, Median values tend to bob Up-and-Down for K > 200 (not shown) while the variability in LRC estimates increases monotonically. All of this suggests that roughly K = 50 is "optimal."

## Systematic Sensitivity Analyses

How sensitive is the location and shape of the distribution of Local Effect-Size estimates to different clustering parameter-settings? To answer this question, analysts using LC Strategy [5] must first decide which of the available X-confounder characteristics are included, and which are excluded, in the process of clustering experimental units. All potential Y-outcome measures (the "left-hand side" variables in model-fitting approaches) as well as the primary treatment indicator or exposure-level measure must be excluded from consideration in the LC process of forming clusters.

The analyst also controls selection of clustering algorithm. The default algorithm [5] is *ward.D* but *diana* or *complete* linkage are also available. However, *single* linkage is *not* recommended!

## Prediction of Local Effect-Sizes

LC strategy *deliberately* separates *estimation* of "local" effect-sizes via (nonparametric) *unsupervised* learning from their *supervised prediction*. Once random-permutation testing *confirms* that the X-confounders used in clustering are *Highly Unlikely* to be ignorable, *prediction* of local effect-sizes using *supervised* methods can be attempted with reduced concern about being misled by overfitting.

Furthermore, when only a few of the most relevant X-confounders are used in clustering, the *within-cluster* variability in these X-confounders is essentially minimized. In this way, key *between-cluster* X-variation is preserved for use in *predicting* the corresponding variation in both y-outcomes and in treatment / exposure relationships. My experience is that *better predictions* are then much more likely to result!

In this optional final phase of LC strategy, the "stretch goal" is to reveal the extent to which local effect-size estimates are heterogeneous across clusters. In other words, the objective is to show that LRC estimates are predictable *fixed* effects rather than homogeneous (unpredictable) *random* effects.

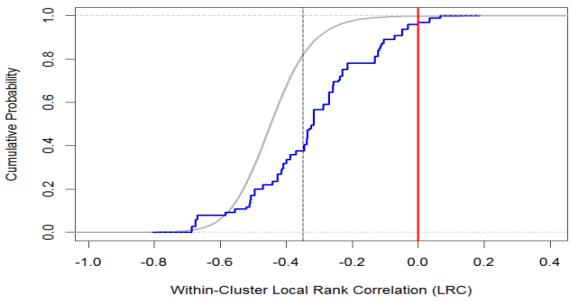


**Figure 3.** Two empirical Cumulative Distributions Functions for LRC estimates are shown here. The eCDF with 50 steps depicts the observed LRC distribution, while the "purely random" eCDF formed using 1,000 independent replications (each using 50 clusters of the same sizes as the observed clusters) looks quite smooth. The observed K-S D-statistic [4] of +0.454 at roughly LRC = –0.35 (vertical line) seems "gigantic." However, it's p-value cannot be "looked-up" in some standard table because both distributions are *discrete* (contain thousands of tied values.) A valid p-value can only be simulated using *another* 1,000 independent replications. Here, the simulated p-value is less than 0.001 because the largest of 1,000 simulated NULL "D"-values is < 0.22. This clearly suggests that the 3 X-confounders used to form the LRC distribution for 50 clusters (Figure 1) are Not Ignorable and, in fact, should be meaningful predictors of LRC variation across US Counties.

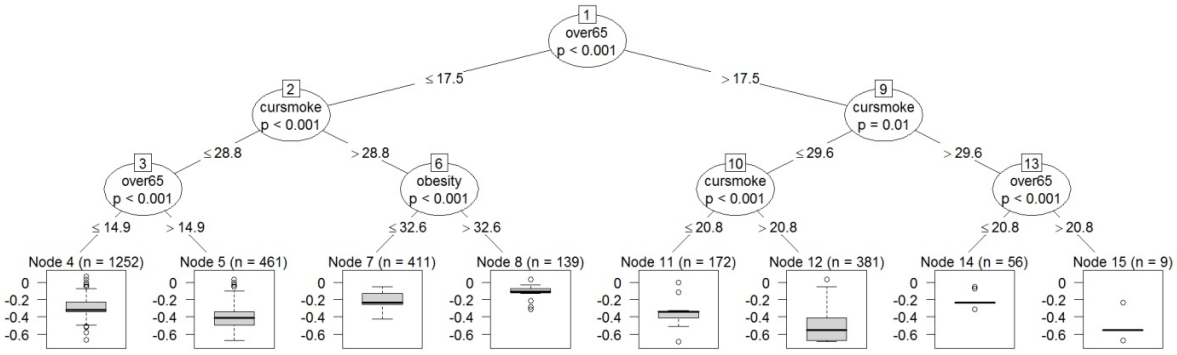


Figure 4. In this small *party tree predictive* model [17,18], final node 4 is quite large (1,252 of 2,881 US counties), and its local distribution of LRC estimates is much like the full distribution for all 2,881 counties. All 7 node splits shown have p-values not only < 0.001 but also less than 0.00015. This ultra-simple model ($R^2$= 0.472) explains slightly less than half of the total across-cluster variation in LRC estimates. These seven splits thus represent Heterogeneous Effects that provide new insights regarding potentially causal relationships detailed in [4].

Because clusters commonly vary considerably in size, it is essential to attach weights to individual local effect-size estimates when fitting traditional "parametric" across-cluster models. Our experience is that simply using weights directly proportional to cluster sizes is both realistic and robust. All available predictor variables, including radon exposure level itself, can then be used in attempts to predict the observed distribution of effect-sizes.

In Figure 4, we illustrate that use of Recursive Partitioning (nonparametric supervised learning) can avoid weighting issues by again relying upon within-cluster tied estimates.

## Final Remark

I hope some readers now feel motivated to run the *demo(pci15k)* or *demo(radon)* examples of LC Strategy [5]. Readers could also modify the R-code within either of the above demos or in my LC Vignette [7] to gain new insights into their own "almost" large cross-sectional datasets.

**References:**


1. Obenchain RL. "Identifying Meaningful Patient Subgroups via Clustering - Sensitivity Graphics." *Biopharmaceutical Section*: *Proceedings of JSM,* 2006.
2. Obenchain RL. "The local control approach using JMP®." Chapter 7 of: *Analysis of observational health care data using SAS*, ed. D. Faries, A. Leon, J. Maria-Haro, R. Obenchain, 151–192, Cary, NC, SAS Press, 2010.
3. Obenchain RL, Young SS. "Advancing statistical thinking in health care research." *Journal of Statistical Theory and Practice,* 2013; 7, 456-469.
4. Obenchain RL, Young SS, and Krstic G. "Low-level radon exposure and lung cancer mortality." *Regulatory Toxicology and Pharmacology*, 2019. doi: https://doi.org/10.1016/j.yrtph.2019.104418
5. Obenchain RL. *LocalControlStrategy*: R-package for robust analysis of cross-sectional data. version 1.3.2 https://CRAN.R-project.org/package=LocalControlStrategy, 2019.
6. Obenchain RL. "Are X-confounders Ignorable? ...a Permutation Test using Random Patient Clusters" *MBSW*. 2019. link to presentation.
7. Obenchain RL. *LocalControlStrategy*: R-package Vignette. 2019. http://localcontrolstatistics.org/other/LCstrategy_in_R.pdf
8. Cochran WG. "The effectiveness of adjustment by subclassification in removing bias in observational studies." *Biometrics* 1968; 24, 205-213.
9. Rubin DB. "Bias reduction using Mahalanobis metric matching." *Biometrics* 1980; 36, 293-298.
10. Rosenbaum PR, Rubin RB. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 1983; 70, 41-55.
11. McClellan M, McNeil BJ, Newhouse JP. "Does More Intensive Treatment of Myocardial Infarction in the Elderly Reduce Mortality? Analysis Using Instrumental Variables." *JAMA* 1994; 272, 859-866.
12. Lunceford JK, Davidian M. "Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study." *Statistics in Medicine* 2004; 23: 2937-2960.
13. Bang H, Robins JM. "Doubly robust estimation in missing data and causal inference models." *Biometrics* 2005; 61: 962-973.
14. Ho DE, Imai K, King G, Stuart EA. "Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference." *Political Analysis* 2007; 15, 199-236.
15. Stuart EA. "Matching Methods for Causal Inference: A Review and a Look Forward." *Statistical Science* 2010; 25, 1–21.
16. Standard Form CMS-R-0235L. "Instructions for Completing the Limited Dataset Data Use Agreement, Section 8.a." http://www.cms.gov/ Medicare/CMS-Forms/CMS-Forms, 2012.
17. Hothorn T, Hornik K, Zeileis A. Unbiased recursive partitioning: a conditional inference framework. *J. Comput. Graph. Stat.* 2006; 15 (3), 651–674.
18. Hothorn T, Hornik K, Zeileis A. *party*: A Laboratory for Recursive Partytioning. R-Vignette, 2006.