# Developing Statistical "Twins" for Qualitative Case Study Site Selection Use

Zhiwei Zhang[1] and Fritz Scheuren[1]
[1]NORC at the University of Chicago, 4350 East-West Hwy, Bethesda, Maryland, 20814

## Abstract

Site selections for qualitative substance abuse evaluation studies are usually constrained due to cost concerns not only by the limited number of sites selected but also by the casual and subjective method for selecting particular sites. A methodology aimed at identifying twin-counties with similar demographics yet different substance use and mental health profiles is developed. Social distance matrices are constructed to list the degrees of similarity among all possible pairs of counties within each state based on pre-determined source variables. The distance matrices are obtained separately through socio-demographic characteristics and through the substance abuse, mental health and service coverage measures. Two ranking indexes are created and composite scales for ranking are established. The methodology is further assessed elsewhere with qualitative case study result from multi-paired twin sites selected here.

**Key Words: survey methodology, site selection, case study, similarity, dissimilarity, distance matrix, Kish, statistical twins**

## Study Design Considerations

Sponsored by the Federal Appalachian Regional Commission (ARC), NORC at the University of Chicago has conducted a study on the Disparities in Substance Abuse, Mental Health, and Access to Treatment in the Appalachian region. The project has two parts: one is quantitative which involved analyses of tens of thousands of household respondents' self-reported data, millions of community hospital patient records, and more than ten thousand specialty treatment facilities' information; the second one, relatively small, is qualitative or what we called the case studies. The objectives of the case study were twofold: To develop a set of criteria and protocols to identify relevant case study communities and conduct case study analyses accordingly; and, to determine the extent of local assessments of the mental health and substance abuse situations as well as the perceived validity of nationally available quantitative data.

As of early 2007, there were 410 counties in the Appalachian region from a total of 13 states. The unit of the site was determined as a county. Cost constrains led to an initial common understanding among stakeholders that only a handful of counties will be selected as sites for focus group interviews. What is presented here is specifically on the site selection considerations, the procedures, and the selection outcomes.

A statistical as well as empirical question was how these sites should be selected with scientific merit and cost beneficial efficiencies. Whereas only a few sites were envisioned to be selected, the substantive areas would have three different yet related aspects – substance abuse, mental health, and access to treatment. We proposed and planned to select 3-6 sites for case studies as part of the investigation.

We aim to reduce to cost by not using too many sites in this exploratory qualitative case study. "Controls are needed in research design for two essential reasons. One concerns the efficiency and economy of research projects and this may be stated in terms of reducing ("minimizing") either the variances or the costs of the projects. The second reason concerns the biases arising from disturbing variables in nonrandomizied designs"(Kish, 1987: 96).

Table 1: Potential Strategies in Research for Controls by Selection versus Analysis

| For Controls by Selection | For Controls by Weighting in Analysis |
|---|---|
| 1. Categorical data | Continuous, linear, normal variables |
| 2. Experimental designs | Survey sampling |
| 3. Few, simple statistics | Complex, multipurpose analysis |
| 4. Data on whole population | Reweighting of sampled units |

Source: Kish, 1987.

Our intention was to select through control – a method echoing Kish's monumental work[1] (i.e., Goodman and Kish, 1950; Kish, 1987; Hess and Heeringa, 2002). Considering the small number of sites involved and that the foremost purpose of the case study part of the project was not to infer the findings to other sites but, rather, to gain in-depth understanding of the substantive matter, the study design would treat the case study site selection as neither a routine sampling issue nor a typical experimental design.

## Sample Survey Theory and/or Experimental Design in Case Study Site Selection

Theory of sample survey chiefly provides estimates of means and totals which are not the goals of case study. Experimental design is used primarily in the analytic search for relationships but case study offers responses to open questions instead of the quantified "y" variable. For a research situation such as a case study, neither true experiments nor conventional sample surveys are practical. It is not a surprise that case study site selections have frequently been governed by judgment samples.

We use, however, experimental design and sampling principles as guidance: From the experimental design perspective (Kish, 1987), we would minimize biases arising from disturbing variables if nonrandomized design is used. From the sample theory, it is desirable to reduce either the variances or the costs of the projects to achieve project efficiency and economy. Of course, this is well documented in numerous existing literatures (Kish, 1965). Our site selection design may run in parallel with ideas that have motivated specific designs such as the "balanced sample survey design." (e.g., Liu and Scheuren 2003). Regardless whether case studies are exploratory or confirmatory in

[1] "Control by selection may seem rather simple when planning for only a few and only relatively simple statistics. But for more complicated and more numerous statistical results, additional controls for disturbing variables would be difficult. Nevertheless, control by selection may prove to be wise oversight, especially for multipurpose surveys." (Kish, 1987: 98)

nature, the sites of the case studies should be physically in existence all the time, with high probabilities that information from extant sources are available and can be utilized, such as the case in this study.

## County Site Selection Design

Our site selection design is “county-centered”. The foremost and obvious reason is that the Appalachian region shaped by West Virginia and 12 other states has been defined by a cluster of geographically adjacent counties. In addition, Counties are in and of themselves important administrative jurisdictions. Many substance abuse and mental health prevention and treatment programs are set up and instituted at the county level. Uniform sub-county data are extremely rare and examining sub-county data is not practical.

The principal of experimental design would point to three things in selecting units: remove the effects of lurking variables on the responses; use randomization to assign subjects to treatments; replicate to reduce chances of strange occurrences affecting the results. How can we remove the effects of lurking variables? If we make the conditions shaping the environments of the sites identical, we would remove outside influences. We used socioeconomic characteristics of the counties as a way to measure the county equalities. Therefore, the selection approach can be designed to target matched counties. To reduce variance, the selection can be further stratified by state. In order to gain further understandings of the disparities of substance use and mental health disorders (SAMH) in this region, we decided to select counties with diverse, or even better –polarized, SAMH statuses.

## Motivation and Utility of Paired and “Twin” Counties

Constructing pairs of counties within states and identifying “twins” are key features of this site selection design. The idea was motivated by the fact that counties –as natural or existing settings – cannot be randomly assigned one type or level of substance abuse and mental health problems at one time and another type or level at another time. To drastically reduce or eliminate noises caused by cofounding variables, we can use pairs of counties as subjects, allowing the “twin” pair of counties to serve as its own control, as long as the appropriate “treatment” and “control” units can be set up.

## Procedures

Within each of the Appalachian state, all possible pairs of counties are listed as sampling units. Each of these paired units are measured and characterized by its socioeconomic status and the substance abuse, mental health, and treatment access statuses.

We construct ranking profiles of the county pairs in terms their socio-demographic characteristics and substance abuse and mental health characteristics and make selections of pairs of counties as field sites for case studies. In essence, three steps are involved: first, pertinent source measures are identified and retrieved as the source variables;

second, statistical procedure is performed to calculate and set up the distance matrices for all Appalachian counties within each state and the distance matrices are transformed into pairs which are subsequently ranked and sorted based on the distance values; third, selection criteria are set up and the final pairs are selected.

### Measures (I): County-level Socio-demographic Characteristics.

The following county-level measures of socio-demographic characteristics are selected as the bases upon which to compare the similarities among counties. These measures are from three major sources – the Appalachian Regional Commission (ARC), the Area Resource File (ARF), and the National Survey on Drug Use and Health (NSDUH).

a. The 2003 *population size* estimates are from the 7/1/2003 County Population Estimates File for Internet Display from the U.S. Bureau of the Census.
b. The *2000 population density per square mile* estimates are from the 2000 Census.
c. The *2000 percentage of urban population* is from the 2000 Census.
d. The *2003 Urban Influence Codes (UIC)* divide the counties, county equivalents, and the independent cities in the United States into 12 groups based on population and commuting data from the 2000 Census of Population, in the case of Metropolitan counties, and adjacency to metro area in the case of non-metropolitan counties[2].
e. The *2000 median home value* is from the 2000 census.
f. The *2004 economic development level codes* are provided by the Appalachian Regional Commission.

### Measurement (II): County-level Substance Abuse, Mental Health, and Service Delivery Statuses

The selected measures[3] and their original sources are listed in the following:

a. *Alcohol abuse or dependence in past year* is from the 2002-2004 pooled National Survey on Drug Use and Health.
b. *Abuse or dependence of any illicit drugs in past year* is from the 2002-2004 pooled National Survey on Drug Use and Health.
c. *Non-prescription use of painkillers in past year* is from the 2002-2004 pooled National Survey on Drug Use and Health.
d. The *percentage of persons having serious psychological distress problems in past year* is from the 2002-2004 pooled National Survey on Drug Use and Health.

[2] The codes were originally from the U.S. Department of Agriculture's Economic Research Service (ERS) website http://www.ers.usda.gov/Data/UrbanInfluenceCodes/

[3] More measures were considered, including: cigarettes use, binge drinking, past month marijuana use, perceptions of risks of drinking and smoking from household surveys. After preliminary statistic analysis to identify patterns of variations (i.e., via factor analysis), these variables were dropped from being used to construct the similarity matrices.

e. The *percentage of persons in correctional or juvenile institutions in past year* is calculated using measures from the Area Resource File
f. The *percentage of persons in mental health hospitals or institutions* is calculated using measures from the Area Resource File
g. The *suicide rate* is calculated using the average numbers of suicides in the past three years and population size from the Area Resource File.
h. An *index on the Health Professional Shortage Area (HPSA) status* is created based on two measures -- the 2003 codes for HPSA for Primary Medical Care[4] and the 2003 codes for HPSA for Mental Health[5], both were originally from the Bureau of Primary Health Care (BPHC) and are available in the Area Resource File.

## Identifying "Twin" Counties through Measuring Similarities of County-Pairs

Another key feature of our site selection design is identifying "twin" counties through measuring the level of similarities of county pairs. The technical approach is to construct statistical distance matrices to list the extent of similarities among all the possible pairs of counties within each state. Operationally, we measure quantitatively through the PROC DISTANCE procedure in SAS.

Existing or prior information are used and sought for the design. We consider a variety of major pertinent variables as the input base variables for the Distance calculations. In particular, the input variables are grouped separately, surrounding socio-demographic characteristics and problems on substance abuse, mental health, and access to treatment. Most of the socio-demographic variables were from the Census bureau. The economic development level codes were provided by the Appalachian Regional Commission. The urban influence codes used definition originated from the Department of Agriculture's Economic Research Services.

## Measuring the Similarities of County-Pairs

Proximity measures provided in the DISTANCE procedure accept four levels of measurement: nominal, ordinal, interval, and ratio. Ordinal variables are transformed to interval variables before processing. This is done by replacing the data with their rank scores, and by assuming that the classes of an ordinal variable are spaced equally along the interval scale.To address the potential issue that variables with large variances tend to have more effect than those with small variances, the input variables with different measurement levels (interval, ordinal) have been taken into account through standardization before the similarity measures are computed.

We rank the all the possible pairs of the counties. This was done through several steps:
Step 1: create county pair distance matrices using socio-demographic measures.
Step 2: create county pair distance matrices using substance abuse and mental health measures, including the access to treatment measures.

[4] http://bhpr.hrsa.gov/Shortage/hpsacritpcm.htm
[5] http://bhpr.hrsa.gov/Shortage/hpsacritmental.htm

Step 3: perform standardizations. We did this to address the potential issue that variables with large variance tend to have larger effect than those with small variances; we also took into account the different measurement levels (such as interval ad ordinal measures).

Step 4: in a state-specific matrix, both the rows and columns are the same counties. We then transform all the state-specific matrices into data rectangular data structure, so that downstream calculations on the final rankings can be made. In this file, the unit is a county pair rather than a county.

The distance matrixes are obtained separately through socio-demographic characteristic and through the substance abuse, mental health and service coverage measures. As a result of this procedure, two ranking indexes are created, namely, soc_rank and samh_rank, indicating the socio-demographic and substance abuse and mental health related similarities separately. The lower the value of the ranking index, the more similar the pair of the two counties are.

## Setting Up Composite Ranking Index

We create separately ranking indexes for socio-demographics and for SAMH characteristics by sorting and ranking the county pairs based on the distance estimations.

Composite for pair of county i and county j in state k is created using:

$$\textit{Ranking Index}_{ijk} = \textit{SAMH Ranking Index Score}_{ijk} - \textit{SOC Ranking Index Score}_{ijk,}$$

The resulting value is used to rank pairs of counties in such a way that the higher value on the composite ranking index would indicate greater dissimilarity on substance abuse and mental health related measures and greater similarity on socio-demographic characteristics.

## Selecting One Pair of Counties from Each State, Selecting Three Backup Pairs of Counties from Each State

One pair of counties, along with three alternative pairs as backup pairs, is selected from the top of the list. The short list of county-pairs selected at the final stage from the 11 Appalachian States is in Table 1.

Considering possible measurement error, through the composite ranking score, three or four pairs of counties from each of the Appalachian states are selected as the candidates of case study sites. To balance bias reduction and efficiency, we evaluate different options and decided to conduct a total of six case studies with three pairs of counties.
If the states were regarded as the strata, certain states were selected with certainty. We made the selection exclusively on the central and the lower part of the northern Appalachian region which had the highest SAMH problems overall.

Three statistical twins from Kentucky, Virginia, and West Virginia, with each pair having the highest composite ranking scale score in the corresponding state, are selected as the final sites for the case studies. These counties are: Morgan county and Wayne county from Kentucky; Bath county and Bland county from Virginia; Hardy county and Monroe county from West Virginia

## Future Applications

Case study design can be implemented in an adaptive way and may be improved with refined measures. The statistical "twins" methodology developed here can be extended to sites other than counties such as universities, business firms, social clubs, etc. depending on the study focuses.

## Acknowledgement

The statistically twinned rankings and the twinned county site selections were performed at NORC at the University of Chicago. The results were shared and consulted with members of Coalition on Appalachian Substance Abuse Policy (CASAP) and conferred with a research team at Eastern Tennessee State University which executed the actual case studies. This was part of one component[6] of a larger study on *Disparities in Substance Abuse, Mental Health, and Access to Treatment in the Appalachian Region*.

## References

Goodman, R. and Kish, L. (1950) Controlled Selection – A Technique in Probability Sampling. *Journal of the American Statistical Association*, 45, 350-372.

Gower, J. C., and Legendre, P. (1986) "Metric and Euclidean Properties of Dissimilarity Coefficients." *Journal of Classification*, 3, 5–48.

Hess, I. and Heeringa, S.G. (2002) *Controlled Selection Continued*. Survey Research Center, Institute for Social Research. Ann Arbor, MI: The University of Michigan

Kaufman, L. and Rousseeuw, P. J. (1990) *Finding Groups in Data.* New York: John Wiley & Sons.

Kish, L. (1965) *Survey Sampling.* New York: John Wiley & Sons.

Kish, L. (1987) *Statistical Design for Research*. New York: John Wiley & Sons.

Liu, Y. and Scheuren, F. (1999) Median Balanced Sampling Design. Proceedings of Survey Research Method Section. Alexandria, VA: American Statistical Association.

[6] The quantitative component includes analyses of 22,000 Appalachian household respondents, 500,000 Appalachian admissions to substance abuse treatment, 167,000 community hospital inpatient discharges, and 980 Appalachian treatment facilities. The qualitative component includes several case studies.

**Table 1. Selected Pairs of Counties Per State in the Appalachian Region, Based on the Composite Ranking Scores**

| County Pairs | | Distance | | Index Rank | | Composite Rank |
|---|---|---|---|---|---|---|
| **County 1** | **County 2** | **Soc-demo** | **SAMH** | **Soc-demo** | **SAMH** | |
| **Alabama** | | | | | | |
| Tallapoosa | Talladega | 0.09399 | 0.63828 | 58 | 662 | 604 |
| Talladega | Marshall | 0.10589 | 0.60212 | 75 | 651 | 576 |
| Lawrence | Chilton | 0.07074 | 0.53362 | 25 | 575 | 550 |
| **Georgia** | | | | | | |
| Stephens | Chattooga | 0.086763 | 0.60025 | 76 | 654 | 578 |
| Jackson | Gilmer | 0.061062 | 0.46431 | 34 | 579 | 545 |
| Jackson | Fannin | 0.091484 | 0.52175 | 86 | 617 | 531 |
| **Kentucky** | | | | | | |
| **Wayne** | **Morgan** | **0.041546** | **0.52521** | **23** | **1206** | **1183** |
| Morgan | Monroe | 0.059565 | 0.50804 | 42 | 1172 | 1130 |
| Morgan | Adair | 0.037319 | 0.48774 | 19 | 1115 | 1096 |
| **Mississippi** | | | | | | |
| Montgomery | Chickasaw | 0.09058 | 0.64941 | 11 | 266 | 255 |
| Winston | Montgomery | 0.08435 | 0.49241 | 10 | 245 | 235 |
| Noxubee | Montgomery | 0.12209 | 0.4788 | 22 | 242 | 220 |
| Winston | Tippah | 0.12187 | 0.46412 | 21 | 240 | 219 |
| **New York** | | | | | | |
| Chautauqua | Allegany | 0.56845 | 0.15715 | 84 | 14 | 70 |
| Tioga | Steuben | 0.19801 | 0.60495 | 21 | 87 | 66 |
| Cattaraugus | Allegany | 0.41395 | 0.0828 | 68 | 4 | 64 |
| Tioga | Cattaraugus | 0.17875 | 0.53048 | 16 | 73 | 57 |
| **North Carolina** | | | | | | |
| Surry | Rutherford | 0.04658 | 0.65497 | 10 | 386 | 376 |
| Yadkin | Madison | 0.05974 | 0.64618 | 14 | 377 | 363 |
| Davie | Alexander | 0.06306 | 0.62176 | 16 | 366 | 350 |
| Surry | McDowell | 0.10259 | 0.67956 | 48 | 395 | 347 |
| **Ohio** | | | | | | |
| Morgan | Meigs | 0.09381 | 0.41246 | 25 | 367 | 342 |
| Noble | Monroe | 0.09847 | 0.41375 | 28 | 368 | 340 |
| Washington | Hocking | 0.11086 | 0.40282 | 39 | 359 | 320 |
| Ross | Hocking | 0.11931 | 0.41563 | 50 | 369 | 319 |
| **Pennsylvania** | | | | | | |
| Somerset | Crawford | 0.024427 | 0.44986 | 14 | 1294 | 1280 |
| Snyder | Juniata | 0.045775 | 0.48158 | 39 | 1311 | 1272 |
| Somerset | Bradford | 0.01667 | 0.42379 | 6 | 1263 | 1257 |
| Huntingdon | Crawford | 0.016052 | 0.41279 | 5 | 1246 | 1241 |
| **Tennessee** | | | | | | |
| Franklin | Claiborne | 0.055416 | 0.58012 | 32 | 1123 | 1091 |
| Overton | Morgan | 0.066463 | 0.58349 | 56 | 1130 | 1074 |
| Scott | Grundy | 0.070694 | 0.58661 | 64 | 1136 | 1072 |
| Roane | Putnam | 0.070714 | 0.58204 | 65 | 1126 | 1061 |
| **Virginia** | | | | | | |
| **Bland** | **Bath** | **0.06587** | **0.78649** | **14** | **259** | **245** |
| Highland | Bland | 0.08249 | 0.72867 | 17 | 256 | 239 |
| Highland | Floyd | 0.0928 | 0.65025 | 20 | 240 | 220 |
| Floyd | Bath | 0.07618 | 0.61837 | 15 | 231 | 216 |
| **West Virginia** | | | | | | |
| **Monroe** | **Hardy** | **0.043199** | **0.60952** | **23** | **1472** | **1449** |
| Pendleton | Monroe | 0.054856 | 0.61336 | 35 | 1475 | 1440 |
| Lewis | Barbour | 0.043253 | 0.54555 | 24 | 1416 | 1392 |
| Wyoming | Barbour | 0.045604 | 0.53389 | 26 | 1402 | 1376 |

Note: County pairs in bold font in this table contain the selected counties which are used subsequently in actual cases studies.