Available online at www.isas.org.in/jisas

JOURNAL OF THE INDIAN SOCIETY OF AGRICULTURAL STATISTICS 64(2) 2010 209-218

# Design, Implementation, and Analytical Methods for a Countywide West Nile Virus Seroprevalence Survey

**Christopher Kippes[1*] and Joseph Sedransk[2]**

*[1]Cuyahoga County Board of Health, Cleveland, Ohio, USA*
*[2]Department of Statistics, Case Western Reserve University, Cleveland, Ohio, USA*

Received 13 January 2010; Revised 18 March 2010; Accepted 20 March 2010

**SUMMARY**

During 2002 there were 221 confirmed or probable cases of West Nile Virus (WNV) in Cuyahoga County (located in Northeast Ohio) – accounting for 71% of all Ohio cases. In December 2002, the public health community of Cuyahoga County conducted a household-based seroprevalence survey designed to estimate focal and county-wide WNV infection rates. In this article the authors provide a detailed description of the field operations used to conduct a countywide serologic survey for WNV and describe the methodology used to obtain and analyze a probability-based sample of households. Field operations were based on incident command structure (ICS) resulting in the recruitment of over 1,200 eligible participants. Although ICS has not been routinely incorporated into traditional public health investigations, it was successfully implemented in this survey. Additionally, the sampling design used in this survey may be helpful in situations where the characteristic of interest has a small, variable probability of occurrence and it is desired to find a large number of individuals with this characteristic - to permit one to relate local rates to local conditions.

*Keywords*: Cluster sampling, Seroprevalence, Stratification, West Nile virus.

## 1. INTRODUCTION

From 2000 through 2003, West Nile Virus (WNV) spread rapidly from east to west across the United States and Canada. Areas such as Louisiana, Illinois, Michigan, and Ohio saw epidemic activity during the spring and summer of 2002.

During 2002, Ohio reported 310 cases of WNV encephalitis and/or meningitis (i.e., West Nile neuroinvasive disease [WNND]), an incidence of 28 cases per million, with 31 fatalities (1). This was the third highest number of human WNND cases in the country, exceeded only by Michigan (557 cases, 56 WNND cases/million) (2) and Illinois (553 cases, 44 WNND cases/million) (3).

The outbreak experienced in Cuyahoga County (which is located in northeast Ohio and includes Cleveland) consisted of 221 confirmed or probable cases of WNV illnesses (including cases of WNV fever) resulting in 11 disease-associated deaths. Of the 221 cases 155 (seventy percent) were WNND cases (incidence, 112 cases/million) (1).

In comparison to previously studied outbreaks in the United States, the 2002 WNV outbreak in Cuyahoga County appeared to be marked by a significant increase in the incidence of passively identified human cases and to affect a wider age range of the population. In December 2002, at the close of this first season of WNV disease transmission, the public health community of Cuyahoga County conducted a household-based seroprevalence survey designed to estimate focal and county-wide WNV infection rates, and to identify host and environmental factors associated with risk for human infection (4).

[*] *Corresponding author* : C. Kippes
*E-mail address* : ckippes@ccbh.net

In this article the authors provide a detailed description of the field operations used to conduct a countywide serologic survey for WNV and describe the methodology used to obtain and analyze a probability-based sample of households in Cuyahoga County. Some features of these field operations may be useful for other surveys that must be conducted under strict budgetary and time constraints. The sample design we used may be useful in circumstances when the characteristic of interest has a small, variable probability of occurrence and it is desired to find a large number of individuals with this characteristic – to permit one to relate local rates to local socioeconomic and environmental conditions.

We have identified other published papers that describe serologic surveys for WNV (5-7). However, there are limitations in each of the survey designs. One survey (5) sampled individuals at outpatient clinics or health centers. This permitted blood samples to be obtained easily, but did not provide a probability-based sample of all individuals in the area under study. A survey in Queens, New York (6) used a probability-based (cluster sample) design, but the area sampled was limited to the epicenter of the WNV outbreak. Similarly, a stratified cluster sample of residents was selected from an area in southern Connecticut (population about 99,000) chosen because of its high crow mortality rate (7).

By contrast, we have carried out a probability-based sample in a large U.S. county. This enabled us to estimate infection rates in areas not near the center of the WNV outbreak. Thus we can relate local infection rates to local environmental and socio-economic characteristics. For example, we can see whether there were local areas with no apparent WNV activity (e.g., no previously identified WNV cases) that actually have significant infection rates. If there are such areas we can investigate the reasons for this apparent anomaly (e.g., the individuals with WNV in an area with low socio-economic status may not seek medical assistance).

## 2. FIELD OPERATIONS

Strict budgetary and time constraints required that the sample be carried out within a two week period. To increase the chance of finding potential respondents at home most of the interviewing took place in the evenings on weekdays and during the days on weekends.

### Training

The Cuyahoga County Board of Health staff decided to create an incident command structure (ICS) that outlined critical components necessary to implement the serosurvey. Historically, military and safety forces have used ICS when responding to major emergencies. As seen in Fig. 1, there are 5 major components to ICS: command, logistics, planning, finance, and operations. This structure serves as a tool for command, control, and coordination of activities. It delineates roles and responsibilities for a given event. To facilitate the development of the key roles and responsibilities, protocols were obtained from the individuals involved with the serosurvey conducted in Queens, New York City in 1999. In addition to creating the ICS structure, a formal training program was developed to standardize the field operations.

### Community Notification

After selection of the clusters (groups of city blocks) to be sampled (see the section, Sample Design, for details), but before the implementation of the survey, city officials were notified of their selection for the survey. Additionally, residents living within the sampled clusters were notified with a flyer one to two days prior to the survey in that area. A local newspaper reporter accompanied the survey team in the field and several television reporters also covered the survey. This provided additional publicity which was expected to increase the participation rate.

### Field Logistics

The survey sample consisted of 88 clusters each consisting of approximately 50 households. Each cluster had a group leader who supervised three task forces (a.k.a. teams), each consisting of a team leader, at least two interviewers and a phlebotomist. Each cluster map had three routes, randomly assigned to the teams. The goal of each team was to obtain at least one blood sample from each of three separate households along their assigned route. To obtain a sample from a tenth household in the cluster, the first team to complete its initial assignment sampled an additional household.

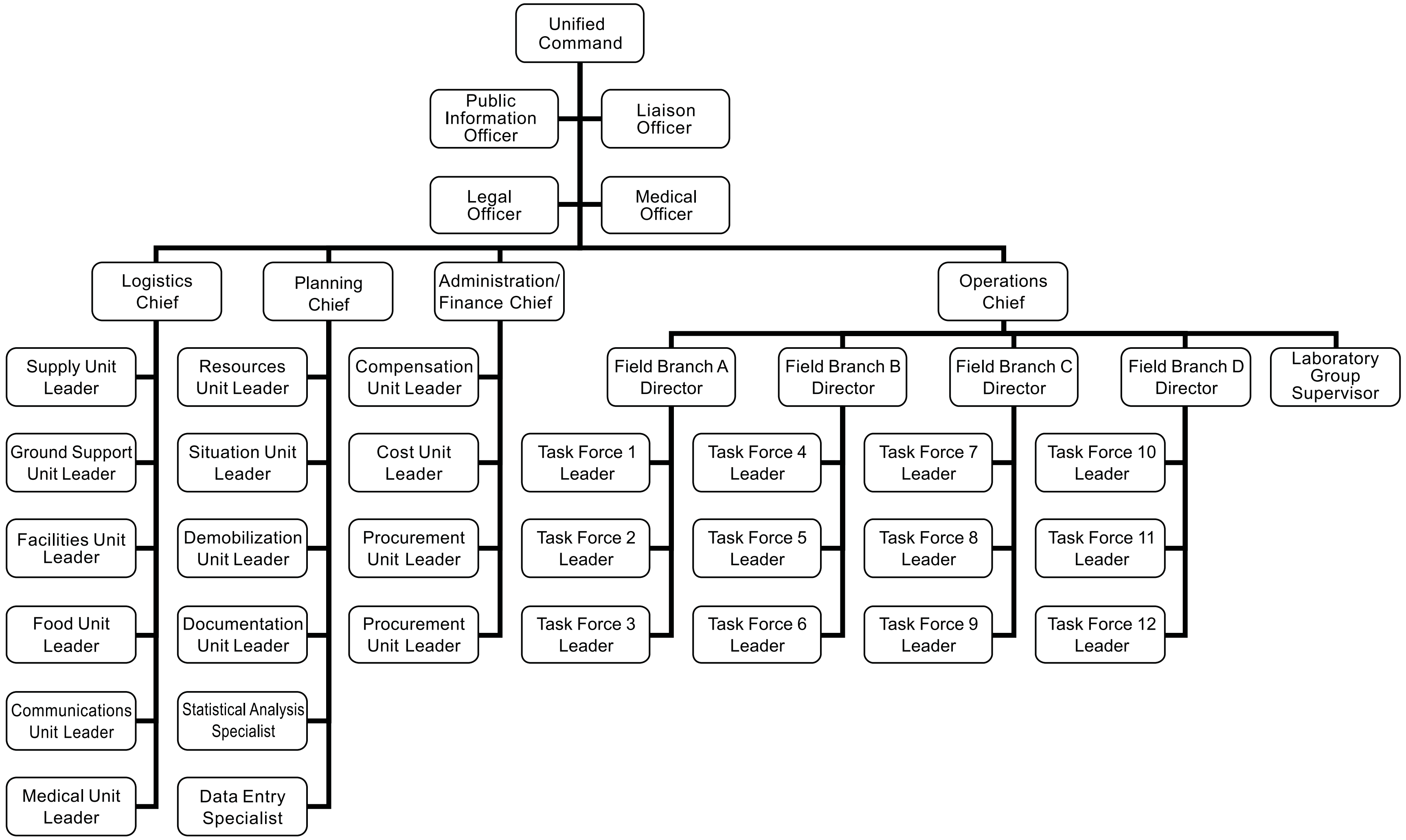

**Fig. 1.** West Nile Virus Seroprevalence Survey Incident Command Structure (ICS)

### Enrollment

Each selected household was approached by a group or team leader, and identification was provided to the head of the household. If at least one adult member of the household agreed to participate, the field staff entered the home and gave an informed consent form to each person who was present. Each child less than 18 years of age required a parent's or guardian's signature on the consent form. In addition, assent was obtained from all children aged 8 and older. Institutional Review Board approval was obtained from University Hospitals of Cleveland.

## 3. SAMPLE DESIGN

Our objective was to select a probability-based sample of households in Cuyahoga County to provide (a) good estimates of the West Nile virus prevalence rate for Cuyahoga County and subpopulations of special interest, and (b) to obtain enough positive cases to permit investigation of the relationship between the presence/absence of West Nile virus and socioeconomic and environmental factors. With limited resources for sampling, a cluster design was needed.

Two covariates were used to categorize the sampled areas: (a) the number of (passively identified) confirmed/probable West Nile virus cases per 1,000 population in 2002 (CRATE), and (b) the mosquito infection rate (IRT) which was defined as the number of positive pools per 1,000 mosquitoes tested, the latter being available for 84 of the 94 municipalities (this includes Cleveland's neighborhoods, and we refer to these neighborhoods as municipalities in the sequel). We started by plotting the IRT against CRATE (Fig. 2) for all municipalities in Cuyahoga County. Clearly, the four municipalities in the upper right corner of figure 2 (Brook Park, Jefferson, Old Brooklyn, Parma Heights) are the areas having the greatest a priori evidence of West Nile virus. The Kinsman area (no confirmed cases, but an infection rate > 45) is also an outlier.

We assigned to Stratum 1 the municipalities of Brook Park, Jefferson, Old Brooklyn and Parma Heights, the presumed epicenter of West Nile virus in Cuyahoga County. As such these areas were of special interest. Moreover, we hoped that by sampling in each of them we would achieve our second objective, (b), described above.

To obtain valid prevalence estimates for Cuyahoga County, sampling of areas with no confirmed cases was needed. Sampling in this stratum, labeled Stratum 3, will also provide direct evidence about the relationship

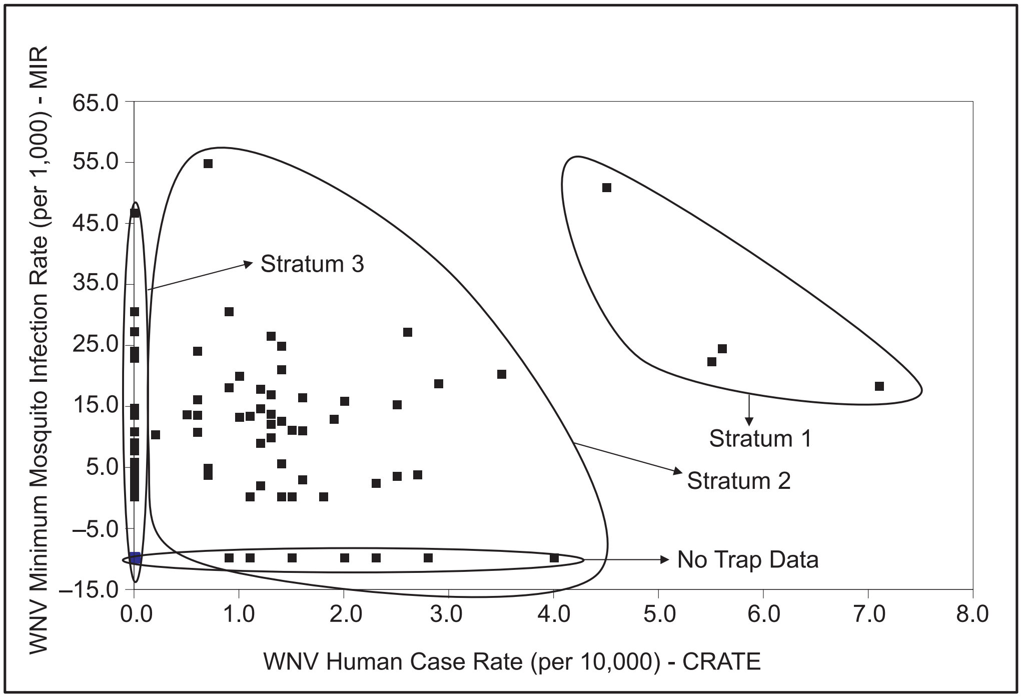

**Fig. 2.** Municipality/neighborhood level mosquito infection rates and human case rates for West Nile virus by strata.

between passively identified and actual/sub-clinical cases of WNV. Because of its special characteristics we selected Kinsman with certainty and then chose two other municipalities with probability proportional to the number of households in each municipality (Buckeye-Shaker and Solon were selected).

For the remaining stratum, Stratum 2, we first divided Cuyahoga County into two sectors, East and West, to ensure coverage of both parts of the county. Within each of the two geographical regions we sampled census tracts rather than municipalities. Of course, these two populations of census tracts excluded the tracts in municipalities belonging to Strata 1 and 3.

The next step of the sample design was to sample census tracts in each of the seven municipalities in strata 1 and 3 and in each of the two regions of Stratum 2. Within each sampled census tract we further sampled two blocks with probability proportional to the number of households in the block. Finally, we sampled about ten households within each sampled block using a procedure that approximated simple random sampling. The blocks were constructed, using 2000 U.S. census data, to form contiguous areas containing approximately 50 households (exception: blocks consisting of multi-family structures contained more than 50 households).

To select the sample of census tracts within each municipality or region, we used the two covariates, CRATE and IRT. For Stratum 2 we divided the census tracts in each region into four substrata (CRATE=0, IRT < 10 or not available; CRATE = 0, IRT > 10; 0<CRATE<.45; CRATE>.45). Within each substratum we sampled two tracts with probability proportional to the number of households in the tract (using the systematic probability proportional to size (pps) method (8, pp. 265-266). Note that the average IRT for Cuyahoga County was 10, and we used scatterplots of CRATE vs. IRT for the tracts to choose the boundary, .45, for CRATE. We chose four census tracts from each municipality in Stratum 1, using the systematic pps method within each of two strata. To facilitate the second objective (i.e., to observe as many cases as possible) we chose three tracts from among those with CRATE > .45 and one tract from those with CRATE < .45).

The municipalities in Stratum 3 had no confirmed cases, so we stratified the tracts in each one as (a) positive IRT and (b) IRT = 0 or no mosquito trap data. In each municipality we sampled the one member of (a) with certainty and three others from (b) with probability proportional to their number of households.

## 4. ESTIMATION

Because there are several survey objectives and a limited budget the sample design, described above, is more complicated than the customary household survey. Consequently, standard survey software could not capture (for variance estimation) important aspects of the sample design, and it was necessary to program the formulas described in this section of this paper.

### Point Estimation

To estimate the finite population total for a variable of interest we use the unbiased Horvitz-Thompson estimator (8, p. 259)

$$\hat{Y}_{HT} = \sum_{i=1}^{n} (y_i/\pi_i)$$

where $y_i$ is the value of the variable $Y$ for the $i$-th unit selected in the sample and $\pi_i$ is the probability that the $i$-th unit is sampled. In our study this probability can be determined as follows. First, consider any one of the thirteen geographical areas described in the section, Sample Design: Brook Park, Jefferson, Kinsman, Old Brooklyn, Parma Heights and the eight groups in Stratum 2 defined by location (east or west), IRT and CRATE. For each area the probability that the $t$-th person in household $k$, block $b$, census tract $i$ is selected is

$$\Pr(\text{person } t \text{ selected} \mid k, b, i) \Pr(\text{household } k \text{ selected} \mid b, i) \Pr(\text{block } b \text{ selected} \mid i) \Pr(\text{census tract } i \text{ selected}) \quad (1)$$

Let $M_{ibk}$ denote the number of persons in household $k$, block $b$, tract $i$ of whom $m_{ibk}$ are assumed to have been selected using simple random sampling. Also define by $T_{ib}$ the number of households in block $b$, tract $i$ of which $t_{ib}$ are assumed to have been chosen using simple random sampling. Finally, let $L_i$ denote the number of blocks in tract $i$ of which 2 are sampled. Then

$$\Pr(\text{person } t \text{ selected} \mid k, b, i) = m_{ibk}/M_{ibk} \quad (2)$$

$$\Pr(\text{household } k \text{ selected} \mid b, i) = t_{ib}/T_{ib} \quad (3)$$

$$\Pr(\text{block } b \text{ selected} \mid i) = 2T_{ib} \Big/ \sum_{b=1}^{L_i} T_{ib} \quad (4)$$

In each area the tracts were stratified further. Within each substratum some tracts were selected with certainty. Then, for such tracts,

$$\Pr(\text{tract } i \text{ selected}) = 1 \qquad (5a)$$

For the rest of the selections in that substratum

$$\Pr(\text{tract } i \text{ selected}) = K\left(\sum_{b=1}^{L_i} T_{ib}\right)\Big/S \qquad (5b)$$

where $S$ is the total number of households in the census tracts in the substratum excluding those in tracts selected with certainty, and $K$ is the number of (non-certainty) tracts sampled in the substratum.

Thus, the selection probability is given by formulas (1) – (5).

For Stratum 3 there is an additional component of (1) for the two municipalities, Buckeye-Shaker and Solon, which were not selected with certainty. The probability that Solon is selected is

$$\frac{2(\text{population of Solon})}{\{\text{population of Stratum 3 excluding Kinsman}\}} \qquad (6)$$

with an analogous expression for Buckeye-Shaker. Thus, the probability of selection of a person in Solon is the product of the terms in (2) through (6).

Estimation of a population ratio is done by using the ratio of estimated totals.

## Variance Estimation

Variance estimation is complicated because there are four stages of sampling with unequal probabilities of selection at several stages. (As noted in the sample design section, this was done to try to accrue as many cases as possible while also having a clustered, probability-based design.) As the basis for our estimation of the variance of a finite population total we use Theorem 11.2 of Cochran's book (8, pp. 301-302). To provide a workable expression for the variance estimator we have made several approximations: (a) assuming, as is customary, with replacement probability proportional to size (pps) sampling instead of the systematic without replacement pps sampling we actually employed, (b) ignoring the stratification of census tracts within their primary sampling unit (e.g., municipality), and (c) aggregating the eight substrata in Stratum 2 to the four listed in Table 1.

**Table 1.** Composition of strata used for variance estimation

| Stratum | Area |
|---|---|
| 1 | Brook Park |
| | Jefferson |
| | Old Brooklyn |
| | Parma Heights |
| 2* | East, CRATE† = 0 |
| | East, CRATE > 0 |
| | West, CRATE = 0 |
| | West, CRATE > 0 |
| 3 | Buckeye-Shaker |
| | Kinsman |
| | Solon |

*Stratum 2 was defined by dividing Cuyahoga County into two sectors, East and West. Within each of the two geographical regions we sampled census tracts rather than municipalities. These two populations of census tracts excluded the tracts in municipalities belonging to Strata 1 and 3.

†CRATE defined as the number of confirmed/probable West Nile virus cases per 1000 population in 2002 .

The same variance formula is used to estimate the variability due to sampling census tracts, blocks, households and persons in each of the eleven areas listed in Table 1. This estimated variance is the sum of formulas (A.2) and (A.3) in the Appendix.

The estimated variance of an estimated total for Cuyahoga County is the sum of the (a) estimated variances corresponding to the four municipalities in Stratum 1, Kinsman in Stratum 3 (because it was selected with certainty) and the four substrata in Stratum 2 (see Table 1), and (b) the remaining contribution from Stratum 3. (The summing is appropriate because sampling is independent in these entities.) Stratum 3 is different from the other two strata because there is an additional sampling stage; i.e., the initial selection of the municipalities. The contribution of Stratum 3 to the total variance is given by (A.4).

The formula for the estimated variance of an estimated ratio is more complicated. We describe in the Appendix how to obtain the estimated variance of an estimated odds ratio. Then we present a method to obtain an approximate 100(1- $\alpha$) percent confidence interval for the population odds ratio.

## 5. RESULTS

A total of 1,747 eligible residents were approached for the survey. Consent was obtained from 1,251 residents (71.6 percent) representing 819 households. The main results of the survey are presented elsewhere (4).

Overall, we estimated that 1.9% of the county's population became infected (95% confidence interval 0.8 - 4.6%). For stratum 1, the epicenter of the West Nile virus outbreak, we estimated that 2.5% became infected (95% confidence interval 0.6 - 9.2%). Surprisingly, for Stratum 3, the areas with no known passively identified cases, the estimated per cent was 3.3. However, the 95% confidence interval (0.4 - 23.9%) for this stratum indicates uncertainty about the magnitude of the prevalence. Finally, for Stratum 2, the estimated infection rate was 1.5% (95% confidence interval 0.2 - 4.4%), lower, as expected, than that in Stratum 1.

## 6. DISCUSSION

In this paper we describe the field operations used to conduct a seroprevalence survey. Although ICS has not been routinely incorporated into traditional public health investigations, it was successfully implemented in this survey. By creating a formal structure for the operation component of the survey, we were able to manage a large number of workers over a large geographical area. Additionally, incorporating ICS into investigational methods will allow public health professionals to gain experience with the system, useful if these individuals need to work closely with safety and fire professionals.

In addition to detailing field operations, this paper provides statistical methodology that may be useful for sample surveys with some of the following features: 1) the characteristic of interest has a small probability of occurrence; 2) it is desired to estimate the prevalence of this characteristic over a large geographical area, including places where activity is expected to be high and places where activity is expected to be low; 3) an objective is to find a (relatively) large number of individuals having the desired characteristic – to permit one to relate local rates to local social and environmental conditions; and/or 4) the survey must be conducted within a very short time period with a limited budget.

For future studies of this type we recommend having a longer time period for the conduct of the survey. We were precluded from doing this because of a limited budget and the onset of the holiday season. Continuing the survey would have increased the likelihood of families not being at home and further extended the time between potential exposure to WNV, the recollection of summer time activities, and the completion of the survey questionnaire. Having an extended survey period would permit a strictly probability-based selection of households with an opportunity to make several callbacks to reduce the nonresponse due to families who are not at home. Estimating the effect of households whose members refuse to participate is difficult. However, selecting a small sample of initial refusals and then using alternative methods to obtain cooperation may provide a useful estimate of the difference between the two groups.

## 7. ABBREVIATIONS USED

| | | |
|---|---|---|
| CDC | : | Centers for Disease Control and Prevention |
| CRATE | : | 2002 West Nile Virus Human Case Rate per 1,000 |
| ICS | : | Incident Command Structure |
| IRT | : | Mosquito infection rate per 1,000 mosquitoes tested |
| PPS | : | Probability proportional to size sampling |
| WNND | : | West Nile neuroinvasive disease |
| WNV | : | West Nile Virus |

## ACKNOWLEDGEMENTS

We thank the residents of Cuyahoga County who participated in this study and applaud the tireless efforts of the public health staff who made the serosurvey a huge success.

## REFERENCES

1. West Nile Virus Update: Human Cases in Ohio, 2002. The Ohio Department of Health (Columbus, OH), (updated Feb 25, 2004; cited Mar 4, 2004). (http://www.odh.state.oh.us/ODHPrograms/ZOODIS/WNV/wnvupdate.htm).
2. 2002 Lab Positive Human West Nile Virus Cases. The State of Michigan Department of Community Health (Lansing, MI), (cited Jan 14, 2004). (http://www.michigan.gov/images/Humanposmap2_74080_7.jpg).
3. 2002 West Nile Virus Surveillance Data: Numbers at a Glance. The Illinois Department of Public Health (Springfield, IL), (updated Oct 1, 2003; cited Jan 14, 2004). (http://www.idph.state.il.us/envhealth/wnvsurveillance02.htm).
4. Mandalakas, A., Kippes, C., Sedransk, J., Kile, J., Garg, A., McLeod, J. *et al*. (2005). West Nile virus epidemic, northeast Ohio, 2002. *Emerging Infectious Disease*, **11(11)**, 1774-1777.
5. Tsai, T.F., Popovici, F.C., Campbell, G.L., *et al*. (1998). West Nile encephalitis epidemic in southeastern Romania. *Lancet*, **352**, 1-5.
6. Mostashari, F., Bunning, M., Kitsutani, P. *et al*. (2001). Epidemic West Nile encephalitis, New York, 1999: Results of a household-based seroepidemiological survey. *Lancet*, **358**, 261-264.
7. Hadler, J., Nelson, R., Lis, M. *et al*. (2001). West Nile virus surveillance in Connecticut in 2000: An intense epizootic without high risk for severe human disease. *Emerging Infectious Disease,* **7(4)**, 636-642.
8. Cochran, W.G. (1977). *Sampling Techniques* (3rd edition). John Wiley, New York.

## APPENDIX

### Variance and Interval Estimation

Let $Y_{ibkl}$ denote the value of the variable of interest for the *l*-th person in household *k*, block *b*, census tract *i*. We suppress the index for the geographical area (e.g., Brook Park; East, CRATE = 0) in Table 1. Assume that there are: (a) $M_{ibk}$ persons in household *k*, block *b*, census tract *i* of whom $m_{ibk}$ are assumed to have been selected using simple random sampling; (b) $T_{ib}$ households in block *b*, tract *i* of which $t_{ib}$ are assumed to have been selected using simple random sampling; (c) $L_i$ blocks in tract *i* of which $l_i = 2$ are sampled independently with probability proportional to $T_{ib} \Big/ \sum_{b=1}^{L_i} T_{ib}$, and (d) *N* tracts of which $n = 4$ are sampled independently with probability

$$z_i = \sum_{b=1}^{L_i} T_{ib} \Big/ \sum_{i=1}^{N} \sum_{b=1}^{L_i} T_{ib} \quad \text{for tract } i$$

Define $$\hat{Y}_{ib} = \frac{T_{ib}}{t_{ib}} \sum_{k=1}^{t_{ib}} \sum_{l=1}^{m_{ibk}} \left( \frac{Y_{ibkl}}{m_{ibk}} \right) M_{ibk}$$

$$\hat{Y}_i = \frac{1}{2} \left( \sum_{b=1}^{L_i} T_{ib} \right) \left( \sum_{b=1}^{2} \frac{\hat{Y}_{ib}}{T_{ib}} \right)$$

and $$\hat{Y} = \frac{1}{4} \sum_{i=1}^{4} \frac{\hat{Y}_i}{z_i} \qquad \text{(A.1)}$$

where the formula in (A.1) is the estimated total for a geographical area in Table 1 under the assumptions outlined in the subsection Variance Estimation.

The two components of the estimated variance of an estimated total are

$$\frac{1}{12} \sum_{i=1}^{4} \left\{ \frac{\hat{Y}_i}{z_i} - \frac{1}{4} \sum_{i=1}^{4} \frac{\hat{Y}_i}{z_i} \right\}^2 \qquad \text{(A.2)}$$

and $$\frac{1}{16} \sum_{i=1}^{4} \frac{\left( \sum_{b=1}^{L_i} T_{ib} \right)^2}{z_i} \left\{ \frac{\hat{Y}_{i1}}{T_{i1}} - \frac{\hat{Y}_{i2}}{T_{i2}} \right\}^2 \qquad \text{(A.3)}$$

The estimated variance of an estimated total, $\hat{v}$, for each of the areas listed in Table 1 is the sum of (A.2) and (A.3).

For Stratum 3 (see Table 1) there is an additional stage of sampling since the municipalities are selected first. (Kinsman is a special case because it was sampled with certainty and, thus, it is treated like the four municipalities in Stratum 1.) Let $\hat{Y}_s$ and $\hat{v}_s$ denote the point estimate (formula (A.1)) and variance estimate (sum of (A.2) and (A.3) ) for Solon with corresponding definitions of $\hat{Y}_b$ and $\hat{v}_b$ for Buckeye-Shaker. Also, define $p_s$ as the population of Solon divided by the population of Stratum 3 except for Kinsman with an analogous definition for $p_b$. Then the estimated variance of an estimated total for Stratum 3 (excluding Kinsman) is

$$\frac{1}{2} \{ (\hat{v}_s / p_s) + (\hat{v}_b / p_b) \} + \frac{1}{2} \left\{ \left( \hat{Y}_s / p_s \right) - \left( \hat{Y}_b / p_b \right) \right\}^2 \qquad \text{(A.4)}$$

Estimation of the variance of an estimated ratio uses a Taylor series approximation. To illustrate, consider the estimated odds ratio $\hat{R} = (\hat{C} + \hat{D})\hat{A} / (\hat{A} + \hat{B})\hat{C}$ where $\hat{A}, \hat{B}, \hat{C}$ and $\hat{D}$ are the estimated total number of persons in Cuyahoga County with (a) positive seroprevalence and positive binary characteristic (e.g., female), (b) negative seroprevalence and positive characteristic, (c) positive seroprevalence and negative characteristic, and (d) negative seroprevalence and characteristic. Let $A_{ibkl} = 1$ if the *l*-th person in household *k*, block *b*, census tract *i* is positive for both seroprevalence and characteristic and $A_{ibkl} = 0$ otherwise, with analogous definitions for $B_{ibkl}$, $C_{ibkl}$ and $D_{ibkl}$.

Define for each individual

$$X_{ibkl} = A_{ibkl} \left\{\hat{B}\,(\hat{C}+\hat{D})/\hat{C}\,(\hat{A}+\hat{B})^2\right\}$$

$$- B_{ibkl} \left\{\hat{A}\,(\hat{C}+\hat{D})/\hat{C}(\hat{A}+\hat{B})^2\right\}$$

$$- C_{ibkl} \left\{\hat{A}\,\hat{D}/\hat{C}^2(\hat{A}+\hat{B})\right\} + D_{ibkl} \left\{\hat{A}/(\hat{A}+\hat{B})\,\hat{C}\right\}$$

and replace $Y_{ibkl}$ in (A.1) – (A.4) with $X_{ibkl}$. The total estimated variance (using $X$ in place of $Y$) is the estimated variance of the estimated odds ratio, $\hat{R}$.

To obtain an approximate 100(1 – $\alpha$) per cent confidence interval for the population odds ratio, $R$, we use a confidence interval for log $R$ and then exponentiate the endpoints to obtain the interval for $R$. The interval for log $R$, based on a standard Taylor series approximation is

$$\log\,(\hat{R}) - z(\alpha/2)\left\{\hat{V}(\hat{R})\right\}^{1/2}/\hat{R} \le \log\,(R)$$

$$\le \log\,(\hat{R}) - z(\alpha/2)\left\{\hat{V}(\hat{R})\right\}^{1/2}/\hat{R}$$

where $\hat{V}(\hat{R})$ is the estimated variance of $\hat{R}$ and $z(\alpha/2)$ is the 100{1 – ($\alpha$/2)} percent point of the standard normal distribution.