2010

Potential applications of randomised graph sampling to invasive species surveillance and monitoring.

Mark J. Ducey
University of New Hampshire, mark.ducey@unh.edu

Kathleen M. O'Brien
Rachel Carson National Wildlife Refuge

IUFRO

SCION

New Zealand Journal of Forestry Science

40 (2010) 161-171

www.scionresearch.com/nzjfs

published on-line:
14/09/2010

Potential applications of Randomised Graph Sampling to invasive species surveillance and monitoring

Mark J. Ducey[1,*] and Kathleen M. O'Brien[2]

[1]*Department of Natural Resources and the Environment, 114 James Hall, Durham, NH 03824 USA*

[2]*US Fish and Wildlife Service, Rachel Carson National Wildlife Refuge, 321 Port Road, Wells, ME 04090*

(Received for publication 3 August 2009; accepted in revised form 5 July 2010)

*corresponding author: mjducey@cisunix.unh.edu

Abstract

Many invasive plants and animals disperse preferentially through linear networks in the landscape, including road networks, riparian corridors, and power transmission lines. Unless the network of interest is small, or the budget for surveillance is large, it may be necessary to draw inferences from a sample rather than a complete census on the network. Desired features of a surveillance system to detect and quantify invasion include: (1) the ability to make unbiased statements about the spatial extent of invasion, the abundance of the invading organism, and the degree of impact; (2) the ability to quantify the uncertainty associated with those statements; (3) the ability to sample by moving within the network in a reasonable fashion, and with little wasted non-measurement time; and (4) the ability to incorporate auxiliary information (such as remotely sensed data, ecological models, or expert opinion) to direct sampling where it will be most fruitful. Randomised graph sampling (RGS) has all of these attributes. The network of interest (such as a road network) is recomposed into a graph, consisting of vertices (such as road intersections) and edges (such as road segments connecting nodes). The vertices and edges are used to construct paths representing reasonable sampling routes through the network; these paths are then sampled, potentially with unequal probability. Randomised graph sampling is unbiased, and the incorporation of auxiliary information can dramatically reduce sample variances. We illustrate RGS using simplified examples, and a survey of *Polygonum cuspidatum* (Siebold & Zucc.) within a high-priority conservation region in southern Maine, USA.

Keywords: Forest biosecurity; sampling theory; graph theory; inventory; Japanese knotweed.

† Based on a paper presented at the IUFRO International Forest Biosecurity Conference, 16-20 March 2009, Rotorua, New Zealand

Introduction

Invasive plants and insects threaten the safe, high-quality, affordable raw materials and sustaining environmental services provided by forested lands (National Research Council (NRC), 2002). However, monitoring and surveillance of invasive species remains a challenging aspect of the management chain. In the United States of America (USA), Executive Order 13112 instructed Federal agencies whose activities impact invasives to develop approaches to "monitor invasive species populations accurately and reliably," but guidance toward meeting that directive is sparse. Accuracy and reliability demand that surveillance and monitoring approaches be statistically sound; realistic budget constraints demand that those activities be efficient and feasible. A report by the Ecological Society of America (Lodge et al., 2006) also highlights the need for efficiency, recognising the high cost of sampling for organisms that may often be rare and

ISSN 0048 - 0134 (print)
ISSN 1179-5395 (on-line)

clustered on the landscape. As Stark et al. (2006) note, the need for, and interest in, risk-based biosurveillance has outstripped the development of sampling methods that are both theoretically and practically attractive.

Unless one has the resources to census an entire population, or an entire region of interest, sampling is fundamental to scientific inference and management decision making. Sampling theory (as a branch of statistics) has always been well-grounded in agricultural and forestry applications, beginning with the early work of Sir R. A. Fisher at Rothamsted, UK (e.g. Fisher, 1925) and the pioneering studies of Hubback (1927). A great deal of sampling in agriculture and forestry is focused on estimating the means (or totals) of attributes in two-dimensional spaces, such as farm fields or forest stands. It is often useful to be able to estimate sampling variances, not only to evaluate the certainty of the results in hand but to be able to design future sampling campaigns and experiments in a cost-effective manner (Gregoire & Valentine, 2008). Within forest biosecurity, sampling theory has progressed to statements of confidence about what is not there: if we fail to detect an incursion, is it because the organism is not there, or because we didn't look hard enough (Coulston et al., 2008)? Whether estimating abundance and impact, or substantiating lack of incursion, a sound, probability-based sampling approach is fundamental to efforts to detect and combat invasive species (Lodge et al., 2006).

Unfortunately, the focus on sampling in two-dimensional regions (such as fields or forest stands) has led to a paucity of tools for sampling in linear networks (such as road networks). Yet surveillance and monitoring in forest biosecurity may often need to focus on road and other infrastructure networks, for three main reasons:

1. some organisms of concern disperse along such networks. For example, many invasive plants preferentially disperse along road networks and power lines, because these are associated with low canopy cover and elevated soil disturbance (e.g. Spellenberg, 1998; Hansen & Clevenger, 2005);

2. anthropogenic mechanisms of dispersal often facilitate the movement of invasive organisms along such networks. For example, the transport of wood-boring insects in contaminated logs and in wooden packing materials occurs along road and rail networks, often leading to long-distance dispersal (NRC, 2002; Chornesky et al., 2005). Mowing and other forms of right-of-way maintenance often serve to propagate invasive organisms that can reproduce from plant fragments (e.g. Oliver, 1996). Soil pathogens (including some key *Phytophthora* species) may be dispersed along trail networks by contaminated boots (Webber & Rose, 2008); and

3. given limited resources for surveillance, it may be more efficient to concentrate effort along transportation networks because potential measurement time is not spent in off-network travel. Depending on the sample objectives and given limited resources, it could be more efficient to constrain the sampling frame and concentrate effort along transportation networks for such species. Surveillance efforts can cover a greater geographic region within a finite budget if measurement focuses on, and respects the opportunities and constraints for field work imposed by, the transportation network.

Furthermore, as we will suggest later in this paper, sampling for invasive arthropods or pathogens within a discrete set of susceptible stands or sites can also be cast advantageously as a network sampling problem.

Recently, we have been involved in the development of a new sampling method called Randomised Graph Sampling (RGS) (Ducey, in press; Knapp & Ducey, 2010). Randomised graph sampling is specifically designed for sampling networks such as road, rail, power line, and trail networks; Knapp and Ducey (2010) presented an application for recreational trail impact assessment. Ducey (in press) has presented mathematical proofs of the statistical attributes of RGS. Randomised graph sampling has several characteristics that are desirable in a sampling method for surveillance and monitoring:

- it allows unbiased estimates of the current status of, and change in, the spatial extent, abundance, and impacts of an invading organism, because it is probability-based;

- it allows unbiased estimates of sampling variance, and provides the ability to quantify uncertainty, again because it is probability-based;

- it is specifically designed to allow efficient sampling while moving through the network, using routes (or "walks") that respect operational constraints. These can include barriers to movement, as well as minimum or maximum time or cost constraints for an operational "piece" of sampling effort (such as a single crew-day); and

- it allows the use of auxiliary information (such as remotely sensed data, ecological models, expert opinion, or volunteer surveys) to focus sampling effort where the organism is most likely to occur, improving sample efficiency and the probability of detecting an incursion. However, it does not sacrifice its probability foundation, and so continues to allow unbiased estimates when such information is used.

The desirable attributes that depend on being a

ISSN 0048 - 0134 (print)
ISSN 1179-5395 (on-line)

probability sampling method are shared by common sampling methods, including simple random sampling and stratified sampling, while other methods (such as importance sampling) also allow the use of auxiliary information. Randomised graph sampling combines these advantages with operational considerations that may be important in practice.

The goal of this paper is to suggest the application of RGS within a biosecurity context, and to outline possible situations where it might be used. Our emphasis will be conceptual and practical rather than mathematical. Unbiasedness and variance concepts will be demonstrated using simple examples rather than proofs (as in Ducey in press). We will illustrate one possible application with a survey of an invasive plant (*Polygonum cuspidatum* Siebold & Zucc.; also *Fallopia japonica* Houtt.) (Japanese knotweed) along roadsides in a high-value conservation area in southern Maine, USA.

Randomised graph sampling

Overview

Randomised graph sampling is a probability-based sampling method originally designed for estimating the parameters of statistical populations associated with linear networks, such as road networks. In mathematical terms, any such network can be described as a "graph": a collection of "edges" (for example, road segments) connecting "vertices" (for example, road intersections). The attributes of interest might be located along the edges (for example, instances of an invasive plant occurring along road segments) or at the vertices (for example, the vertices might be reasonable trap locations for pheromone trapping of a wood-boring insect; the edges would then represent reasonable travel routes between trap locations).

Figure 1 shows a simplified example of a mathematical graph that could represent a road network. In this case, the edges represent road segments between vertices (intersections), and are labelled with attribute values (including the road segment length, and the number of metres of road frontage that are infested with an invasive plant). Rather than sampling individual vertices or edges, in RGS we sample "walks" or feasible sampling routes. (The term "walk" is taken from graph theory in mathematics; in practical situations a survey might be conducted by walking, driving, or any other suitable travel method). For example, if all feasible routes must start and end at vertex A at the bottom of Figure 1, then ABCDBA, ABCDEBA, ABCDECBA, and ABCEBA represent candidate walks. Statistical estimation in RGS allows for the overlap among walks, and for the possibility (and even desirability!) of assigning some walks higher probabilities of selection than others.

[1] *Geographic Information System*

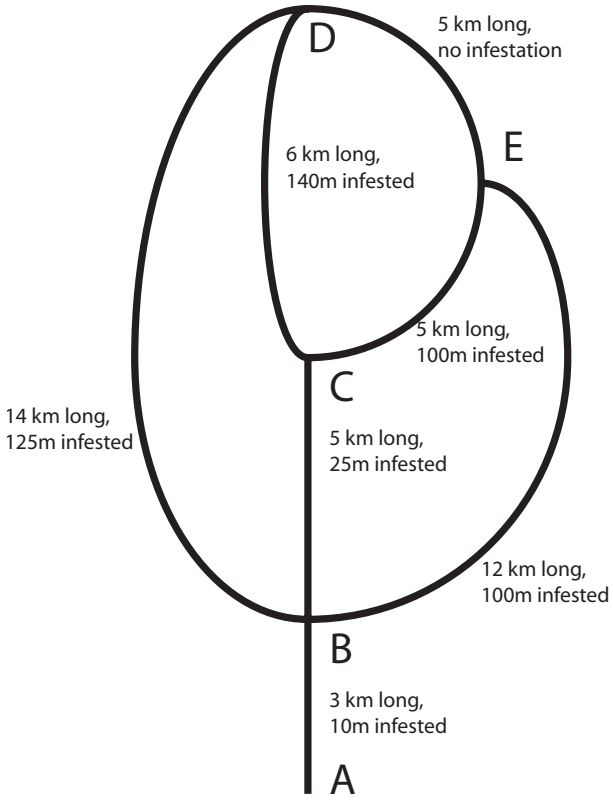

FIGURE 1: An example graph in which attributes (total metres of road frontage infested by an invasive plant) are associated with edges (road segments).

From a practical standpoint, RGS includes the following steps:

I. identify the network of interest. For example, one would obtain a road map (or GIS[1] layer) and label the edges and vertices;

II. identify an adequate set of "walks" or surveying routes. Routes should be constrained to reflect operational considerations (e.g. maximum travel times). The set of routes does not need to be exhaustive. However, each edge and vertex needs to appear in at least one walk;

III. assign sampling probabilities to each walk. In the simplest case, each walk is assigned equal probability (for example, if there are 100 walks, each walk is assigned a 1% chance). However, this is not necessary, and indeed this is where auxiliary information can be employed to great advantage. For example, if interpretation of remotely sensed imagery suggests a concentration of a susceptible forest type along particular walks, those walks can be assigned higher probability than walks that do not appear to include susceptible types. Concentrating sampling effort in areas where the threat is most likely to occur increases efficiency

ISSN 0048 - 0134 (print)
ISSN 1179-5395 (on-line)

and reduces sampling variance; the effect of unequal probability sampling is adjusted using appropriate estimating equations in step V, below;

IV. select walks at random using the assigned probabilities, and conduct the appropriate measurements on each walk; and

V. using the sampling probabilities developed in Step III, and the results obtained in Step (IV), compute the desired estimates using the appropriate equations (Ducey, in press). If multiple walks have been selected, compute the standard error and confidence limits.

Estimating equations

Before proceeding to a more general discussion of steps (I) through (V), and their implementation in a biosecurity context, it will be useful to review the estimating equations for means, totals, and standard errors that must be used in step (V). For the moment, consider the problem of estimating the total of some attribute over the graph based on the selection of a single walk or sampling route. The true total, X, is fixed but unknown to us. We only know the attribute values x_i for the graph elements (edges or vertices) that are actually in the sample. The basic estimator in RGS is a Horvitz and Thompson (1952) estimator,

$$\hat{X}_{RGS} = \sum_i \frac{x_i}{P_i} \qquad [1]$$

where P_i is the probability that the i^{th} element will appear in a randomly selected walk. Let q_j be the probability of selecting the j^{th} walk from the list developed in step (III), using the probabilities chosen in step (IV). (The q_j must conform to the usual rules for probabilities: they must all be greater than zero, and they must sum to one over all the walks in the list.) Let d_{ij} be a simple indicator variable that equals 1 if the ith graph element is included in the j^{th} walk, and equals 0 otherwise. Then it is easy to show that

$$P_i = \sum_j d_{ij} q_j \qquad [2]$$

where the summation is over all the walks in the list. Equation [2] can be calculated for every graph element in advance of sampling, because it does not depend on the x_i. Once a walk has been selected, Equation [1] can be used to provide an unbiased estimate of the total X.

Now suppose we have selected multiple walks, and let $\hat{X}_{RGS,k}$ denote the estimate calculated from the k^{th} walk out of n selected walks. Then the usual sample mean

$$\overline{X} = \frac{1}{n} \sum_k \hat{X}_{RGS,k} \qquad [3]$$

provides the best unbiased estimate of X, and

$$s_{\bar{x}}^2 = \frac{1}{n(n-1)} \sum_k (\hat{X}_{RGS,k} - \overline{X})^2 \qquad [4]$$

provides an unbiased estimate of the squared standard error of $\overline{X}$ if walks were sampled with replacement, or a trivially biased estimate if walk were sampled without replacement and the number of walks in the list is considerably greater than n (as would almost certainly be true in application, unless the graph were very small or the budget for monitoring and surveillance were unusually large).

When the x_i themselves are known imperfectly, either because they are measured on sub-samples of the graph elements or because of imperfect detectability, then Equation [4] will underestimate the uncertainty in $\overline{X}$. However, the probability basis of RGS provides straightforward error propagation in such cases. For example, if graph elements are sub-sampled, then RGS is a variable-probability first stage in a two-stage sampling approach (cf. Thompson, 2002, ch. 13), and should be analysed accordingly.

Assignment of probabilities

Perhaps the most unsettling aspect of RGS for some readers may be the assignment of selection probabilities q_j for walks that occurs in step (III), with direct impact on the inclusion probabilities P_i for individual graph elements. It may seem paradoxical that the probabilities can vary – perhaps even arbitrarily, or based on a subjective assessment of where an organism might be or where change is likely to occur – and yet the resulting estimates will be unbiased.

Unequal probability sampling has a long history in forestry. Perhaps the most familiar example is horizontal point sampling (also known as prism sampling or variable radius plot sampling), in which sample trees are selected with probability proportional to their basal area. Indeed, the practical and theoretical development of horizontal point sampling by Bitterlich (1948) and Grosenbaugh (1958) was contemporaneous with the general theoretical development of unequal probability sampling in the statistical literature (Horvitz & Thompson, 1952). As readers will recall, horizontal point sampling is very efficient for estimating tree volume and biomass because the basal areas of individual trees, and therefore their probabilities of selection, are highly correlated with volume and biomass. In general, Horvitz-Thompson estimators are unbiased no matter what probabilities P_i are used (provided the P_i are all greater than zero) (Thompson, 2002, ch. 6). The variance of a Horvitz-Thompson estimator will be low whenever the ratio x_i/P_i is nearly constant (Horvitz & Thompson, 1952). The price of poor probability assignment is high variance, not bias.

The most obvious choice of q_j is to set them all equal. For example, if one has generated a list of 100 walks

ISSN 0048 - 0134 (print)
ISSN 1179-5395 (on-line)

in step (II), one would set q_j = 0.01 for every walk. This may not result in equal P_i values for all the graph elements, however, as some elements may occur in multiple walks. For example, if all walks for the graph in Figure 1 must start and end at vertex A, then edge AB has a 100% chance of inclusion no matter how many walks are in the list or what their selection probabilities may be. Indeed, equal assignment of the q_j may lead to inclusion probabilities P_i that are poorly correlated (or even negatively correlated!) with the x_i, depending on the structure of the graph and the spatial distribution of the attribute of interest.

To reduce the variance of RGS estimates by improving the correlation between the P_i and the x_i, it will often be helpful to introduce auxiliary information. The simplest case, and one that is an obvious choice when the attributes occur on graph edges as in Figure 1, is to use the edge lengths as auxiliary information. Intuitively, all else being equal, a short graph edge will be likely to contain less of an interesting organism or change than will a long graph edge, and the same is true when those edges are combined into walks. If the length of an edge is l_i, we might assign the selection probability of the j^{th} walk proportional to the total length of its measured edges L_j:

$$L_j = \sum_i d_{ij} l_i \qquad [5]$$

$$q_j = L_j / \sum_j L_j$$

However, all else may not be equal. If we have better information about where an attribute is likely to be found (whether that attribute is the abundance of an organism, or a rate of change), we are free to use that to reduce the variance. "Better information" might come from a sophisticated computer model, from information on possible mechanisms of spread within the graph, from previous surveys (including volunteer surveys that might be incomplete or only partially reliable), from expert opinion, or even crude and subjective information. Whatever the source of the information, suppose we can capture it as a positive number or covariate y_i for each element on the graph, in which the greater the value of y_i the more likely it is that the graph element contains the attribute of interest. Then a simple approach is to set the selection probability of a walk proportional to the sum of its covariate values Y_j:

$$Y_j = \sum_i d_{ij} y_i \qquad [6]$$

$$q_j = Y_j / \sum_j Y_j$$

These simple alternatives are not the only ones available. Ducey (in press) discusses further techniques for optimising the q_j, though these require some mathematics that are beyond the scope of this paper. If there is great uncertainty about whether the available covariates will indeed be correlated with the attributes of interest, then other estimators for multiple walks (such as a generalised ratio estimator; Thompson 2002, pp. 76-79) may provide considerable reduction in variance in exchange for a small amount of bias. Exploration of those alternatives is likewise outside the scope of this overview.

Invasive plant survey

Let us now return to Figure 1, which depicts a highly simplified invasive plant survey scenario. The plant of interest is found primarily along roads, so the road network forms our graph. Alternatively, we may be interested in the road network because roadside rights-of-way are our area of responsibility; we may be surveying roadsides as part of a rapid assessment, perhaps in advance of some more detailed survey to be done later; or we may simply lack the resources or authority to conduct a thorough reconnaissance of interior areas and are focusing on roads out of necessity. In any case, the roads are the domain of our sampling and our inference. Randomised graph sampling strategies when roads and interior areas are of interest will be discussed below.

Obtaining a map of the road network, and labelling the map so that edges and vertices are identifiable, completes step (I) in RGS. Step (II) is to develop a list of walks or feasible sampling routes. Every edge or vertex must occur in at least one walk in the list; otherwise, the list may reflect any operational constraints we may wish to impose or desired features we wish to incorporate. Earlier, we suggested that if all walks must begin and end at vertex A, then ABCDBA, ABCDEBA, ABCDECBA, and ABCEBA would represent candidate walks. These four walks also form an adequate list, because every road segment appears in at least one walk. However, we are free to add ABCDCBA to the list if it suits our fancy, and it does (there is an ice cream stand at intersection C that makes a hot afternoon of field work more tolerable for the crew). Our list now includes 5 walks. ABA would certainly be an easy walk to measure, but it does not cover any graph elements that are not included in other walks, and it does not suit our fancy so we exclude it. If including or excluding a walk confers practical or even perceived advantages to those responsible for designing or executing the campaign, and importantly does no other harm, then it is permissible. So long as the list of walks is adequate, it does not affect the unbiasedness of RGS.

We now reach step (III), the assignment of selection probabilities to the individual walks. To illustrate the simplest case, suppose we assign an equal selection probability q_j = 0.2 to each of the 5 walks in our list. By examining the list of walks, we can determine the P_i for each of the edges of the graph. (We ignore the vertices, as the attribute of interest – number of metres of infested roadside – is associated exclusively with the edges.) Edges AB and BC occur in all 5 walks, so

ISSN 0048 - 0134 (print)
ISSN 1179-5395 (on-line)

$P_{AB} = P_{BC} = 1$. By contrast, edge BD occurs in only one walk, so $P_{BD} = 0.2$. The inclusion probabilities for the other edges are $P_{BE} = 0.4$, $P_{CD} = 0.8$, $P_{CE} = 0.4$, and $P_{DE} = 0.4$.

Steps (IV) and (V) are to select walks, perform the measurements, and compute the estimates associated with the selected walks. Because the probabilities are equal, many random or pseudo-random techniques could be used to select walks. The use of a random number generator in a computer would be ideal, but for the sake of illustration suppose we place 5 numbers in a hat, and happen to draw the number 2 which corresponds to the second walk in the list (ABCDEBA). We conduct our field work (with required ice cream stop) and obtain the x_i values for each segment. $\hat{X}_{RGS}$ for this walk can be calculated as:

$$\hat{X}_{RGS} = 10/1.0 + 25/1.0 + 140/0.8 + 0/0.4 + 100/0.4 = 460 \text{ m}$$

The results for all the walks in the list are shown in Table 1. Because the average of the values of $\hat{X}_{RGS}$, weighted by the equal selection probabilities, equals the true total invasion of the graph (500 m), we can see that the RGS estimate is unbiased. The weighted mean squared deviation of $\hat{X}_{RGS}$ from the true value is the variance of an estimate from a single walk, and that translates into a CV of 40% for single-walk estimates when sampling is with equal probability.

Now suppose that instead of equal probability sampling, we had used sampling with probability proportional to the total measured length of each walk in step (III). For example, the measured length of walk ABCDBA is 28 km (we do not double-count edges when they are traversed a second time on the return trip). The measured lengths for the other four walks are 31 km, 24 km, 25 km, and 14 km, respectively, and the total over the five walks is 122 km. The selection probability of the first walk is thus 28/122 = 0.2295. The selection probabilities for the other walks are given in Table 1. Given the selection probabilities of the walks, calculation of the P_i is straightforward. As before, edges AB and BC occur in all 5 walks, so $P_{AB} = P_{BC} = 1$. By contrast, edge BD occurs only in the first walk, so $P_{BD} = 0.2295$. The inclusion probabilities for the other edges are $P_{BE} = 0.4590$, $P_{CD} = 0.7951$, $P_{CE} = 0.4016$, and $P_{DE} = 0.4508$.

Now suppose that once again, we draw walk 2 (ABCDEBA) as our sample. (We will definitely have needed a random number generator to draw the walks with unequal probability.) Now $\hat{X}_{RGS}$ for this walk can be calculated as

$$\hat{X}_{RGS} = 10/1.0 + 25/1.0 + 140/0.7951 + 0/0.4508 + 100/0.4590 = 428.94 \text{ m}$$

The results for all the walks in the list are shown in Table 1. Because the average of the values of $\hat{X}_{RGS}$, weighted by the unequal selection probabilities, equals the true total invasion of the graph (500 m), we can see once again that the RGS estimate is unbiased. The weighted mean squared deviation of $\hat{X}_{RGS}$ from the true value again gives variance of an estimate from a single walk, which translates into a CV of 32% for single-walk estimates. In this case, even the simple expedient of using edge length as the covariate has led to an appreciable reduction in variance.

The example depicted in Figure 1 is highly simplified. Such a simple graph would hardly require subsampling in real life. However, in application real road, trail, rail, and power-line networks can quickly generate a large number of edges which, in feasible combinations, can allow development of a very large number of candidate walks. In practice, a GIS would be a helpful tool in generating the graph, designing feasible walks, evaluating covariates, and assigning probabilities. The unbiasedness of RGS does not, however, depend on the complexity of the graph or the technology used to manage information about it.

TABLE 1: Estimates associated with sampling the graph depicted in Figure 1, when selection probabilities for walk are equal and with probability proportional to length.

Walk	Equal Probability		Probability Proportional to Length	
	q_j	$\hat{X}_{RGS}$	q_j	$\hat{X}_{RGS}$
ABCDBA	0.2	835	0.2295	755.7
ABCDEBA	0.2	460	0.2541	428.9
ABCDECBA	0.2	460	0.1967	460.1
ABCEBA	0.2	535	0.2049	501.8
ABCBA	0.2	210	0.1148	211.1
	$E[\hat{X}_{RGS}]$	500	$E[\hat{X}_{RGS}]$	500
	CV (%)	40.1	CV (%)	32.4

ISSN 0048 - 0134 (print)
ISSN 1179-5395 (on-line)

Surveillance of susceptible habitats

In its original conception, RGS closely followed the first example, in which attributes occur on the edges of the graph, and the graph directly mirrored a physical network such as a road or trail network. However, other sampling situations may be thought of as graph sampling problems, especially if travel costs and feasibility are important practical considerations.

As an example, suppose we are interested in surveillance for a lethal forest pathogen. The pathogen is known to affect a tree species that occurs in a recognisable stand type. However, existing stand maps or the analysis of remotely sensed data may identify far more patches of the susceptible stand type than can be visited in a reasonable field campaign. Furthermore, travel time between stands may be susbtantial. It might be possible, in principle, to draw a simple random sample of the susceptible stands, but visiting that simple random sample would require travelling directly past other, non-sampled stands that could easily have been visited en route. If travel is costly relative to sampling, then the overall cost efficiency of the simple random sample will be low. Randomised graph sampling would allow greater efficiency, by taking advantage of the proximity of stands and combining them into feasible sampling walks that respect operational advantages and constraints.

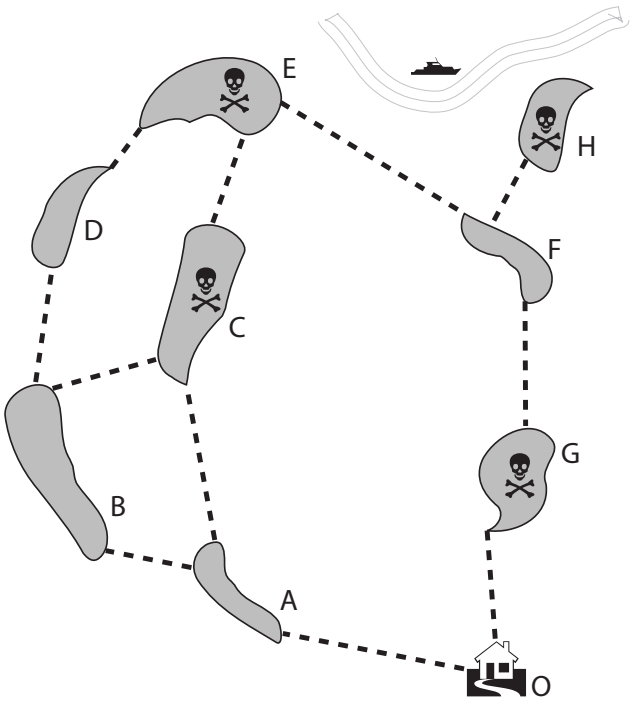

FIGURE 1: An example graph in which attributes (presence or absence of an invasive pathogen, indicated by skull and crossbones) are associated with vertices (patches of suitable forest habitat). Vertex O is the office, where sampling trips must begin and end.

Figure 2 illustrates such a scenario. The field office (vertex O) must be the origin of all feasible walks. In step (I) of RGS, we would use available information to develop a list of susceptible stands; these become the vertices of the graph. The edges of the graph are feasible travel routes connecting nearby stands. Identifying reasonable routes (or at least reasonable travel time requirements), as well as barriers (such as unbridged rivers) that might prevent direct movement between nearby stands, would be an important component of graph identification in step (I).

Once step (I) is complete, we move to step (II), the identification of feasible walks. Suppose that given the travel times and field requirements, it is reasonable to sample 3 stands in a field day; successful completion of 4 stands would be unlikely, while completing only 2 stands would not fully use the available time. Feasible walks might then include OABC, OABD, OACE, OGFE, and OGFH (the return portions are not listed, for simplicity). All vertices appear in at least one walk, so this list is adequate for RGS.

In step (III), we assign probabilities. As in the previous example, we have 5 walks in the list, so we would assign q_j = 0.2 to each walk. In that case, the P_i are just the number of walks that each vertex occurs in, multiplied by 0.2: 0.6 for A, 0.4 for B and C, 0.2 for D, 0.4 for E through G, and 0.2 for H.

In step (IV), we randomly select one or more walks and perform the field work. If we record a 1 for each stand that is infested, and a 0 for each stand that is not, then $\hat{X}_{RGS}$ estimates the total number of stands that are infested. The estimates that would result from the selection of each walk are presented in Table 2. Once again, we find that RGS is unbiased, with an expected or average estimate of 4 infested stands.

Now suppose that in step (III) we had available a new risk map developed by the modeler down the hall. The model predicts that incursions are likely to originate from the port located near stands E and H. It calculates a relative risk for E and H that is 4 times the risk for the farthest stands, A and B. All the other stands, which occur at intermediate distances, are calculated to have a relative risk that is twice that of A and B. If we take the selection probability for a walk as proportional to the sum of its stands' relative risk scores from the model, and proceed accordingly, the resulting probabilities, estimates, and summary statistics will be as presented in Table 2. Once again, RGS is unbiased. Again, the inclusion of covariate information that turned out to be correlated with the attributes of interest did reduce the sampling variance.

ISSN 0048 - 0134 (print)
ISSN 1179-5395 (on-line)

TABLE 2: Estimates associated with sampling the graph depicted in Figure 2, when selection probabilities for walk are equal and with probability proportional to modelled relative risk.

Walk	Equal Probability		Probability Proportional to Modelled Risk	
	q_j	$\hat{X}_{RGS}$	q_j	$\hat{X}_{RGS}$
OABC	0.2	2.5	0.1290	2.82
OABD	0.2	0	0.1290	0.00
OACE	0.2	5	0.2258	4.88
OGFE	0.2	5	0.2581	4.00
OGFH	0.2	7.5	0.2581	5.81
	$E[\hat{X}_{RGS}]$	4	$E[\hat{X}_{RGS}]$	4
	CV (%)	63.7	CV (%)	45.2

Field example

Study site and organism

Practical application of RGS for invasive species inventory can be quite straightforward. As an example, we describe here a rapid survey of *Polygonum cuspidatum* invasion within the Mt. Agamenticus-to-the-Sea conservation area in southern Maine, USA.

The Mt. Agamenticus-to-the-Sea conservation area comprises approximately 4800 ha of forested land in southern coastal Maine, centered on the low peak of Mt. Agamenticus (43.223° N, 70.692° W, elevation 211 m). The area is heavily forested, with dominant tree species including *Pinus strobus* (L.) (eastern white pine), *Quercus rubra* (L.) (red oak), *Acer rubrum* (L.) (red maple), and *Tsuga canadensis* (L. Carr.) (eastern hemlock). Land ownership is a patchwork of local and state government, private conservation organisations, and individual private holdings, with low density residential development and small-scale agriculture as common land uses within a generally forested matrix. The area has experienced nearly four centuries of European settlement, and much of the forest dates from agricultural abandonment in the late 19th and early 20th centuries. The area is one of the most ecologically diverse in the state of Maine, and is home to several locally or globally rare species, including *Emydoidea blandingii* (Holbrook 1838) (Blanding's turtle), *Williamsonia lintneri* (Hagen in Selys 1878) (ringed boghaunter dragonfly), and *Sylvilagus transitionalis* (Bangs 1895) (New England cottontail rabbit), which is a candidate for listing under the US Endangered Species Act. Although sometimes called "the largest unfragmented coastal forest between Acadia National Park and the New Jersey Pine Barrens" (Mt. Agamenticus to the Sea Coalition, 2009), the area is penetrated by nearly 200 km of public roads.

During the early spring of 2009, we undertook a rapid assessment of invasion by *Polygonum cuspidatum* in the northern half of the conservation area. *Polygonum cuspidatum* is an alien invasive capable of forming dense thickets that displace desirable vegetation in pasture and forest systems of the northeastern USA (and in many other regions where it is invasive) (Wade et al., 1996; Forman & Kesseli, 2003; Weston et al., 2005) and can also be invasive in riparian areas. Rhizome and stem fragments are often spread by humans through roadside mowing (Conolly, 1977; Brock et al., 1995), and, in our study region, anecdotal evidence suggests soil disturbance by snowploughing and flooding may also play a role. Once established, Japanese knotweed often forms monospecific thickets that cast deep shade and can exclude native vegetation (Seiger & Merchant, 1997).

Methods and Results

In RGS step (I), we obtained a high-quality map of public roads in the study area, labeling each intersection and each location where a public road passed out of the study area as a vertex. The map included 85 km of roads, describable using 48 segments or edges connecting 44 vertices.

In step (II), we developed a list of feasible sampling walks. Walks were developed by hand using the physical map and knowledge of the study area. Walks were constrained to begin and end at vertices known to have reasonable parking areas nearby, and were required to have a length appropriate to sampling on foot in a 4-6 hour partial field day. We constructed a list containing 41 candidate walks encompassing all the edges and vertices of the map.

In step (III), we assigned sampling probabilities based on *ad hoc* subjective scoring. In general, the study area is at the frontier of *P. cuspidatum* invasion.

ISSN 0048 - 0134 (print)
ISSN 1179-5395 (on-line)

We were already aware of two significant invasions occurring just outside the study area, and a small number of others on frequently travelled roads within the study area. We assigned all segments a base score of 1 point. Segments already believed to contain *P. cuspidatum* were given an additional 16 points, while segments adjacent to those were given an additional 4 points. The resulting point values were multiplied by segment length to arrive at a final score. The final scores were used in Equation [6] to give the selection probabilities for each walk.

In step (IV), we selected 6 walks without replacement, using a spreadsheet to facilitate the selection with variable probabilities. The walks did include some overlapping segments. We surveyed each walk (only surveying the overlapping segments once) and recorded the position and linear extent along the road of each *P. cuspidatum* individual or clump encountered.

In step (V), we used Equations [1] through [4] to obtain estimates associated with each walk, to calculate the mean of those 6 estimates as the best final estimate, and to obtain a standard error for the total number of clumps and total linear extent of clumps within the 85 km road network. The best estimates were 14.0 ± 5.4 clumps, and 105 ± 56 metres of total infestation. Although significant sampling uncertainty remains after surveying only 6 walks, the results suggest current infestation by *P. cuspidatum* is less than had been feared. However, the small size of many clumps, and their distribution among segments, suggests that invasion is actively occurring and may be characterised by relatively long-distance dispersal. This pilot study has provided a useful baseline for a more thorough survey of the entire 200 km road network in the conservation area. We are also planning to conduct a 100% census of the study area reported here, in support of simulation studies that will provide a better understanding of the impacts of more sophisticated sources of auxiliary information on the sampling process.

Discussion and Conclusions

Randomised graph sampling may be useful for a range of survey, monitoring, and surveillance applications in forest biosecurity. However, it is not (and is not intended to be) a universal solution. A strength of RGS is that it respects, and even takes advantage of, constraints on movement through the landscape imposed by existing road or other transportation networks. However, where those constraints do not exist or are not limiting, other approaches will probably be simpler and more effective. For example, if travel is inexpensive and fast and the study area is small, it probably makes sense to select sampling locations as a simple random sample.

The examples presented in this overview are highly simplified, and will not match any particular application in detail. However, they should stimulate further discussion, and to illustrate how the elementary mechanics of RGS can be implemented. Further extensions of the basic RGS framework can capture a variety of sampling situations. For example, in situations where attributes (such as instances of an invasive plant, or infrastructure risks from dead or threatened trees) occur along graph edges (such as roads or power line corridors), it may not be possible or desirable to survey individual edges exhaustively. Randomised graph sampling can be used to select edges as a first stage of sampling, with points, plots, or transects used to sub-sample edges in a second stage. Likewise, if attributes are found in stands located at the vertices of an RGS graph, it may not be possible to conduct a complete census in the selected stands. However, RGS can set a first-stage foundation for a more complex multistage program involving plots for trees, pheromone traps for insects, and so on.

A limitation (and strength) of RGS as developed here is its focus on the graph or network itself as the subject of sampling. Of course, even in cases where a biosecurity threat is found or disperses predominantly along a transportation corridor, interior forests may also be at risk. One possibility is to use RGS within a stratified sampling framework: the landscape can be stratified into areas along transportation corridors, and interior areas. A similar stratification approach has been suggested using other techniques for sampling harvest damage to residual trees along and between skid trails (Stehman & Davis, 1997). The operational efficiency of RGS can be harnessed to sample the more easily accessible and connected areas, while other approaches can be employed for the more expensive and inaccessible interior. Sampling cost should be a consideration in the allocation of effort in stratified sampling (Thompson, 2002, pp. 120-124). Thus, RGS may facilitate the development of more cost-effective surveillance efforts even when transportation networks are not the sole concern of sampling.

Another strength of RGS is that it allows unbiased estimation even though sampling effort is concentrated where risk is greatest. From this perspective, it conforms to emerging trends in targeted sampling for animal biosecurity (Stark et al., 2006; Wells et al., 2009). Coulston et al. (2008) have recently developed freedom-from-infection ideas from the animal biosecurity literature to allow substantiating the freedom from infestation or incursion in a forest biosecurity context. The approaches developed by Coulston et al. (2008) require a valid probability sample as a prerequisite. Extension of their techniques to RGS would be a valuable contribution.

ISSN 0048 - 0134 (print)
ISSN 1179-5395 (on-line)

Acknowledgments

This paper is Scientific Contribution Number 2433 of the New Hampshire Agricultural Experiment Station. Findings and conclusions in this article are those of the authors and do not necessarily represent that of the United States Fish and Wildlife Service.

References

Bitterlich, W. (1948). Die Winkelzählprobe. *Allgemeine Forst- und Holzwirtschaftliche Zeitung 59*(1/2), 4-5.

Brock, J. H., Child, L., de Waal, L. C., & Wade, M. (1995). The invasive nature of *Fallopia japonica* is enhanced by vegetative regeneration from stem tissues. In P. Pyesk, K. Prach, M. Rejmanek, & M. Wade (Eds). *Plant Invasions: General Aspects and Special Problems* (pp. 131-140). Amsterdam, The Netherlands: SPB Academic Publishing.

Chornesky, E. A., Bartuska, A. M., Aplet, G. H., Britton, K. O., Cummings-Carslon, J., Davis, F. W., Eskow, J., Godron, D. R., Gottschalk, K. W., Haack, R.A., Hansen, A. J., Mack, R. N., Rahel, F. J., Shannon, M. A., Wainger, L. A., & Wigley, T. B. (2005). Science priorities for reducing the threat of invasive species to sustainable forestry. *Bioscience, 55*(4), 335-348.

Conolly, A. P. (1977). The distribution and history in the British Isles of some alien species of *Polygonum and Reynoutria. Watsonia, 11*, 291-311.

Coulston, J. W., Koch, F. H., Smith, W. D., & Sapio, F. J. (2008). Invasive forest pest surveillance: survey development and reliability. *Canadian Journal of Forest Research 38*, 2422-2433.

Ducey, M. J. (in press). Randomized graph sampling. *Environmental and Ecological Statistics*.

Fisher, R. A. (1925). *Statistical Methods for Research Workers*. Edinburgh, UK: Oliver and Boyd.

Forman, J., & Kesseli, R. V. (2003). Sexual reproduction in the invasive species *Fallopia japonica* (Polygonaceae). *American Journal of Botany, 90*, 586-592.

Gregoire, T. G., & Valentine, H. T. (2008). *Sampling Strategies for Natural Resources and the Environment*. New York, NY, USA: Chapman & Hall/CRC Applied Environmental Statistics.

Grosenbaugh, L. R. (1958). *Point sampling and line sampling: probability theory, geometric implications, synthesis*. USDA Forest Service, Southern Forest Experiment Station, Misc. Paper 160.

Hansen, M. J., & Clevenger, A. P. (2005). The influence of disturbance and habitat on the presence of non-native plant species along transport corridors. *Biological Conservation, 125*(2), 249-259.

Horvitz, D. G., & Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association, 47*, 663-685.

Hubback, J. A. (1927). Sampling for rice yield in Bihar and Orissa. Bulletin 166, Imperial Agricultural Research Institute, Pusa, India. Reprinted in *Sankhya, A7*, 281-294 (1946).

Knapp, R., & Ducey, M. J. (2010). A cost effective and efficient way to assess trail conditions: a new sampling approach. In S. Weber (Ed.) *Rethinking Protected Areas in a Changing World: Proc. 2009 GWS Biennial Conference on Parks, Protected Areas, and Cultural Sites* (pp. 213-218). Hancock, Michigan, USA: The George Wright Society.

Lodge, D. M., Williams, S., MacIsaac, H. J., Hayes, K. R., Leung, B., Reichard, S., Mack, R. N., Moyle, P. B., Smith, M., Andow, D. A., Carlton, J. T., & McMichael, A. (2006). Biological invasions: recommendations for U.S. policy and management. *Ecological Applications, 16*, 2035-2054.

Mt. Agamenticus to the Sea Coalition. (2009). Retrieved July 31, 2009 from http://www.mtatosea.org/project_area.html.

National Research Council (NRC). (2002). *Predicting invasions of nonindigenous plants and plant pests*. Washington, DC, USA: National Academy Press.

Oliver, J. D. (1996). Mile-a-Minute Weed (*Polygonum perfoliatum* L.), an invasive vine in natural and disturbed sites. *Castanea, 61*(3), 244-251.

Seiger, L. A., & Merchant, H. C. (1997). Mechanical control of Japanese knotweed (*Fallopia japonica* [Houtt.] Ronse Decraene): Effects of cutting regime on rhizomatous reserves. *Natural Areas Journal, 17,* 341-345.

Spellerberg, I. F. (1998). Ecological effects of roads and traffic: a literature review. Global. *Ecology and Biogeography Letters, 7*, 317-333.

Stark, K. D. C., Regula, G., Hernandez, J., Knopf, L., Fuchs, K., Morris, R. S., & Davies, P. (2006). Concepts for risk-based surveillance in the

ISSN 0048 - 0134 (print)
ISSN 1179-5395 (on-line)

field of veterinary medicine and veterinary public health: review of current approaches. *BMC Health Services Research, 6*, 20.

Stehman, S. V., & Davis, C. J. (1997). A practical sampling strategy for estimating residual stand damage. *Canadian Journal of Forest Research, 27*, 1635-1644.

Thompson, S. K. (2002). *Sampling*, 2nd ed. New York, NY, USA: John Wiley & Sons.

Wade, M., Child, L., & Adachi, N. (1996). Japanese knotweed – a cultivated coloniser. *Biological Science Reviews, 8*, 31-33.

Webber, J. F., & Rose, J. (2008). Dissemination of aerial and root infecting *Phytophthoras* by human vectors. In S. J. Frankel, J. T. Kliejunas, & K. M. Palmieri (Tech. Coords.), *Proc. Sudden Oak Death Third Science Symposium* (pp. 195-198). General Technical Report PSW-GTR-214. Washington DC, USA: USDA Forest Service,

Wells, S. J., Ebel, E. D., Williams, M. S., Scott, A. E., Wagner, B. A., & Marshall, K. L. (2009). Use of epidemiological information in targeted surveillance for population inference. *Preventive Veterinary Medicine, 89*, 43-50.

Weston, L. A., Barney, J. N., & DiTommaso, A. (2005). A review of the biology and ecology of three invasive perennials in New York State: Japanese knotweed (*Polygonum cuspidatum*), mugwort (*Artemisia vulgaris*) and pale swallow-wort (*Vincetoxicum rossicum*). *Plant and Soil,* 277, 53-69.

ISSN 0048 - 0134 (print)
ISSN 1179-5395 (on-line)