

Chapter 40: Informed and Targeted Initial Designs

Previously we discussed the standard initial sample designs available in SADA. These designs are classics in the sense that they have been around for many years with an objective of meeting some customary statistical endpoint. Such endpoints include hypothesis testing under assumptions of independent and identically distributed data, searching for objects, or some spatial coverage objective. Most work under the premise (stated or implied) that absolutely nothing is known about the site. This is often not true but adopted in the interest of being conservative. In reality, many environmental data sets are not independent, the costs of evaluation are too high to ignore prior knowledge about contamination events, and the objectives may be highly geographical and entirely outside the realm of traditional statistical objectives.

This chapter discusses initial designs that use prior, spatially definable information to distribute sample locations across the site to meet some target objective. These methods (with the exception of the judgmental design) borrow heavily from the secondary design strategies. In addition, a new method for searching for objects when there is some information about where it most likely exists is introduced.

Sources of Prior Knowledge

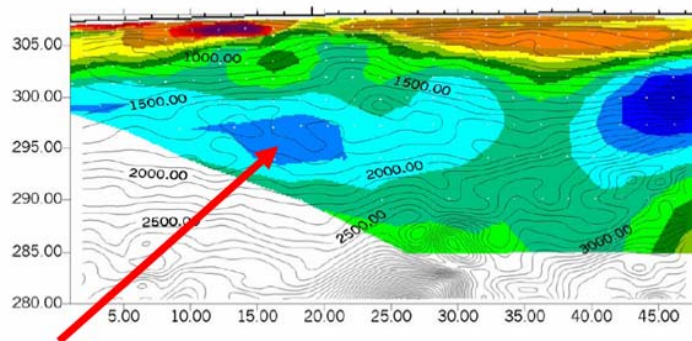
Spatially definable prior information essentially equates to a 2d or 3d model of what site conditions are thought to be prior to any current evaluation. Of course, the issue is how to come up with this model. There are many ways but one must be able to accept a degree of uncertainty about the quality of the model. Here are four example sources:

1) Field detection methods

For surface contamination events, it may be possible to use cheap field detection technologies as a precursor to sampling. Such technologies may include X-Ray Fluorescent (XRF) detectors, various “sniffers”, and radiological scans. Field detection methods that provide less accurate but more abundant data are gaining traction within the regulatory community. For example, the TRIAD model developed and supported by EPA (www.triadcentral.org, last accessed 6/10/2009) encourages this type of increase in data supply over accuracy. The emphasis is shifted to adequate accuracy to support the decision. *Note: SADA is often recognized as a TRIAD support tool (see for example http://www.triadcentral.org/tech/dsp_sub.cfm?id=13, last accessed 6/10/2009).* These detection result can be used to form a model that may indicate the location of contamination in a reasonably rigorous way.

2) Geophysical measurements

In some cases, geophysical measurements may give some insight into the location of subsurface contamination. The following image taken from Watson et al (2002) shows a nitrate plume identified with the use of electrical resistivity (paper can be found at <http://public.ornl.gov/orifc/other/DollSpectrum.pdf>, last accessed 6/9/2009).



Nitrate Plume

(Taken from Watson et al. 2002. Nitrate plume annotation added here.)

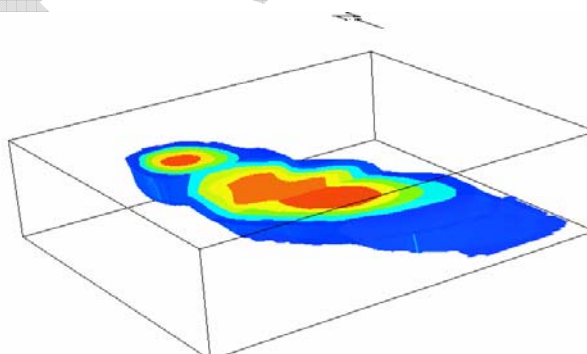
With this kind of geophysical information, it is possible to distribute samples such they have a greater chance of intersecting and even bounding such a subsurface volume.

3) Previous sampling efforts

Some sites may have been under investigation for a longer period of time. Historical data may be available that may remain relevant even after accounting for an extended time lapse. This kind of prior data can be imported into SADA, modeled and used in a secondary or initial targeted design.

4) Professional Expertise (user models)

It may be possible to spatially delineate a “site conceptual” model that indicates where contamination is at based on an assimilation of several lines of evidence. This can include historical documents, previous sampling efforts (that may have lost some relevance), or even some geophysical measurements. In this case, it may be possible to create a user model (see user model chapter) that captures these different lines of information and creates a conceptual endpoint map/volume. The following is a user created delineation of the probability of contamination in the subsurface assimilated from a variety of sources.



If these forms of information can be brought or created in SADA as models the following initial sample designs can be brought to bear.

Targeted High Value Design (simple)

This design places new samples in those locations most likely to be contaminated based on the prior map. This design is exactly the same as the secondary high value design discussed

in the secondary sample design chapter and will not be repeated here. If you are interested, please visit that chapter now. There is no difference between the approaches. If you import a prior model (see chapters on importing modeled or gridded data) then you are in good shape. Simply select the model under the data type imported models, select Develop sample design as the interview and select High Value Design from the list of available designs. If you create the model in SADA using a geospatial model, then the process is exactly the same. To be more efficient you might consider first storing the geospatial model as a static design. Otherwise, SADA will rebuild it each time you build a sample design.

If you create the model in SADA as a user model, then you have a couple of new options: High value design (simple) and simply high value design. In the case of high value design (simple) then SADA treats the user model as if it were a static stored model. If you select “high value design” SADA provides you an opportunity to first update the user model with some real data and then distribute samples.

Area of Concern Design (simple)

This design places new samples in those locations found along the boundary between likely contaminated and likely uncontaminated. This of course implies a decision criteria and a decision framework (see chapter on determining the area of concern). The design is exactly the same as the area of concern boundary design discussed in the secondary sample design chapter and will not be repeated here. If you are interested, please visit that chapter now. There is no difference between the approaches. If you import a prior model (see chapters on importing modeled or gridded data) then you are in good shape. Simply select the model under the data type imported models, select Develop sample design as the interview and select High Value Design from the list of available designs. If you create the model in SADA using a geospatial model, then the process is exactly the same. To be more efficient you might consider first storing the geospatial model as a static design. Otherwise, SADA will rebuild it each time you build a sample design.

Target High Value and Area of Concern Design (non-simple)

In the case of User Created probability models, these designs allow you to do an update of your model using real measured values just before distributing the new samples. Updating user defined models is discussed in the User Models Chapter and won't be repeated here. It is advisable to update your probability map and then store it as a static copy (see previous section). As an imported or stored model you can more efficiently apply the simple versions just discussed. To use this non-simple implementation, you would follow exactly the same parameterization methods discussed in the section on updating the user model in the user model chapter along with exactly the same parameterization techniques discussed in the high value and area of concern designs discussed in the secondary design chapters.

Bayesian Ellipgrid

In standard Ellipgrid applications, such as the hotspot search routines, the underlying assumption is that the elevated zone does exist and the grid will discover it with some prescribed probability. In this sample design, we do not assume that the elevated zone does exist. Instead, the user provides probability coverage for the entire site, indicating spatially where the zone is likely to be. This is done by creating a user defined probability map. SADA then uses this probability map as the basis for a revised ellipgrid approach.

In the ellipgrid world, since $P(E) = 1$ for every ellipsoid, the problem is reduced to a geometric calculation. This same geometric calculation can be used to infer the number and location of samples when prior knowledge is introduced.

The informed ellipgrid approach first accepts the probability map as previously described, along with the standard parameters that describe the size (and possibly shape) of the elevated zone of concern. When the routine is executed, SADA first creates polygons encompassing each zone individually. Within each polygon, the original ellipgrid code already available in SADA may be utilized to consider the probability within each polygon in the following manner.

When including probabilities that an elevated zone exists at all, the discussion becomes more intuitive to consider the probability that an elevated zone exists and is missed rather than the probability that an elevated zone exists and is found. This is particularly true for the search for contaminated zones, and it transitions well into the other SADA discussions which consider the probability that an elevated zone exists given none was discovered by sampling. Therefore, the algorithm is described from this later vantage point.

The original ellipgrid code needs the value for probability of finding the elevated zone. This must be adjusted to account for the probability that it might not be there at all. In particular, we must determine the equivalent probability of discover assuming tht the elevated is definitely there. To accomplish this we begin first with the following definitions.

F = The object is found

DF = The object is not found

E = The object exists

DE = The object does not exist

In reality, an object will exist or not. Our probability of finding it can be completely enumerated as

$$P(F) = P(F|E) \times P(E) + P(F|DE) \times P(DE)$$

This indicates that the probability of finding an object is really a function of our chances of finding it when its definitely there $P(F|E)$ and our chances of finding it when it is definitely not $P(F|DE)$. The latter is obviously zero but we include it in the mathematics to provide a complete enumeration. In order to properly qualify $P(F)$, the probability of finding it given its there $P(F|E)$ must be adjusted by the probability that its there at all. This is done by multiplying $P(F|E)$ by $P(E)$. The same is true for $P(F|DE)$ as well.

As mentioned previously, in the latter term we have $P(F|DE) = 0$. So the equation can be reduced to simply:

$$P(F) = P(F|E) \times P(E)$$

Now $P(F)$ is really the probability parameter the code is expecting. That is, the probability of discovery. If $P(E) = 1$ then we simply have $P(F) = P(F|E)$ and we are in the normal ellipgrid situation.

It is certainly true that to extend the model to account for the probability that it might not be there, we would just need the user to provide $P(E)$. This is certainly possible, but the meaning of $P(F|E)$ is not quite intuitive. This is different now than asking the user to specify the probability of finding it $P(F)$. So for bayesian ellipgrid, we ask What is the probability that we miss it when its really there? In other words what is $P(DF|E)$? This question really turns the question into one of an acceptable risk level for the user. All that is required is to move from $P(DF|E)$, $P(E)$ to $P(F)$.

We have that

$$P(F|E) + P(DF|E) = 1$$

So applying a bit of trivial algebra we have

$$P(F|E) = 1 - P(DF|E)$$

Plugging into the equation above we have

$$P(\text{you find it}) = [1 - P(DF|E)] \times P(\text{exists})$$

So in the Bayesian ellipgrid approach, $P(\text{exists})$ is provided by the user defined prior probability map. The user must also provide the level of risk they are willing to take $P(DF|E)$. The equation above then provides the equivalent $P(F)$ needed for the ellipgrid model.

The standard probability map provides the $P(A)$. Here $P(A)$ is a constant within each polygonal region. The value $P(A,B)$ must be provided by the user. From a technical standpoint, the algorithm calculates the $P(C|A)$ where C is the probability that we detect the elevated zone. It is simple then to calculate this value as

.

This value is automatically calculated by SADA for the user behind the scenes. At this point, there is a constraint imposed by the current version of SADA that says $P(C|A)$ must be greater than 10%. That is, the probability that an elevated zone exists and we missed it cannot be less than 10%.

The remaining parameters specify the size and geometry of the elevated zone and are the same as the standard hotspot search model.

Suppose we have a region that has a probability of containing an elevated zone of only 50/50. Suppose further that we wish the chance of missing such an elevated zone to be no more than 10%. What is the grid size that would accomplish this for a circular zone of radius 100ft?

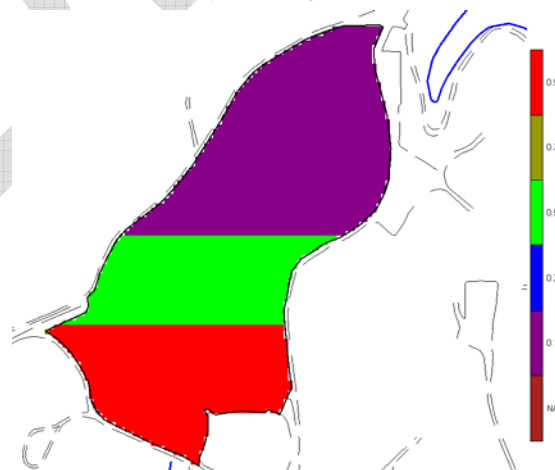
In SADA, we would first create a user defined map that includes the 50/50 region of interest. Then we would choose the bayesian ellipgrid model and indicate a probability value of missing the hotspot equal to 10. SADA then performs any necessary transformations behind the scenes and produces a map of the necessary grid spacing.

Then, for each polygonal area, SADA will calculate the grid required to meet the user need for $P(A,B)$ given $P(A)$. This number will usually be smaller than the standard ellipgrid model, which assumes that the elevated zone does exist. Similarly, areas with smaller probabilities of containing an elevated zone will be searched with less frequency than those with higher probabilities. This makes sense, as intuitively one will spend greater resources in areas of greater potential payoff.

Let's do a quick example.

We recommend that you first read the section on hotspot search designs in the initial design chapters. The parameters and approach there are almost exactly the same as here. Therefore we will cover them here very briefly.

Open the file BayesianEllipgrid.sda. When the file opens you are presented with a probabilistic user model called "Probability Exists". This surface only site has been populated with three value regions (0.15, 0.5, and 0.95) corresponding roughly to "I don't think its here", "I don't know if its here", and "I think its here" respectively.



Select the interview Develop Sample design, click on the set sampling parameters step, and choose Bayesian Ellipgrid from the drop-list of available designs.

Sample Design
Bayesian Ellipgrid
Calculates search grids based on prior knowledge about site conditions.

Hot Spot Search (2d)

Grid Definition
Grid
Square
Length of X side Length of Y side X/Y Ratio

Shape Definition
Hot Spot Shape
Elliptical Shape: 1.0 is a circle 1
Hot Spot Orientation
 Random
 Degrees
Refresh

Hot Spot Definition
 Area of the hot spot 7853.9816339
 Major radius length 50

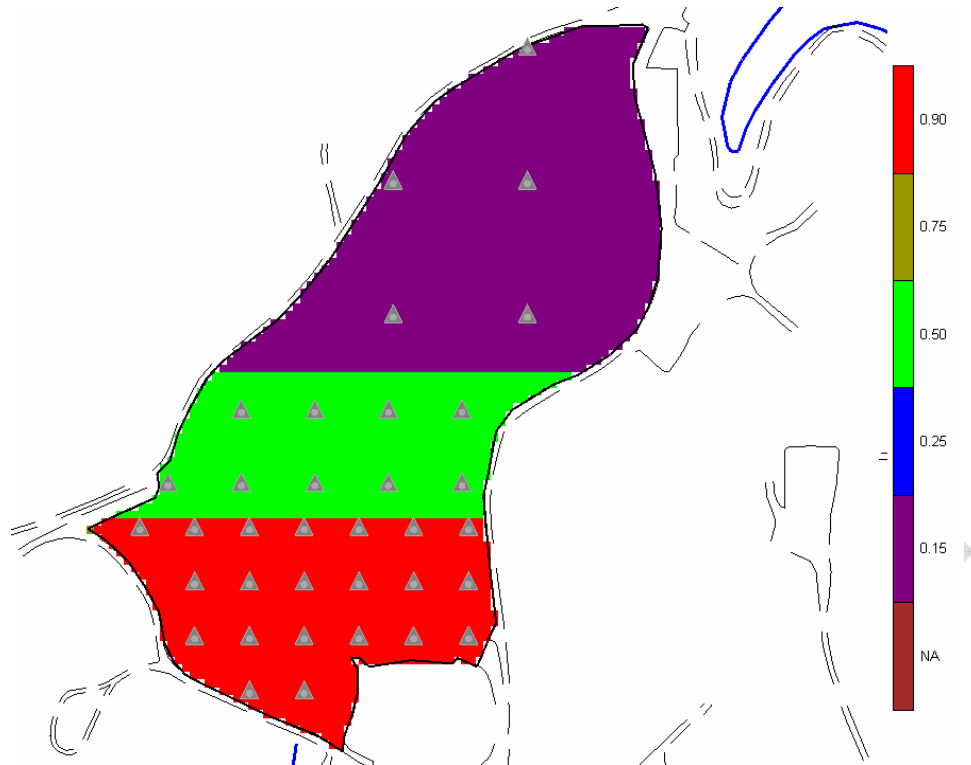
Probability
Probability hotspot exists and we miss it 10 %

We wish to search this site for a circular object with at least a 50ft radius using a square grid. We don't think it exists with equal probability everywhere on the site but when we're done we want the chance that it exists and we simply missed it with sampling to be only 10%. Notice that this is fundamentally different than the probability parameter in standard ellipgrid. There the probability parameter was the probability that we discover it (assuming it does exist).

We begin with the hot spot search (2d) parameter block. Bayesian ellipgrid technically does apply to a 3d model but all layers must be the same model. In a 3d context, you are drawing the probability a little differently. In 3d where all layer models must be the same, you are really saying the chance is X% that it will be found somewhere below this area of the site. Dealing with a sampling model that can also use information about how far down is the subject of a future research effort.

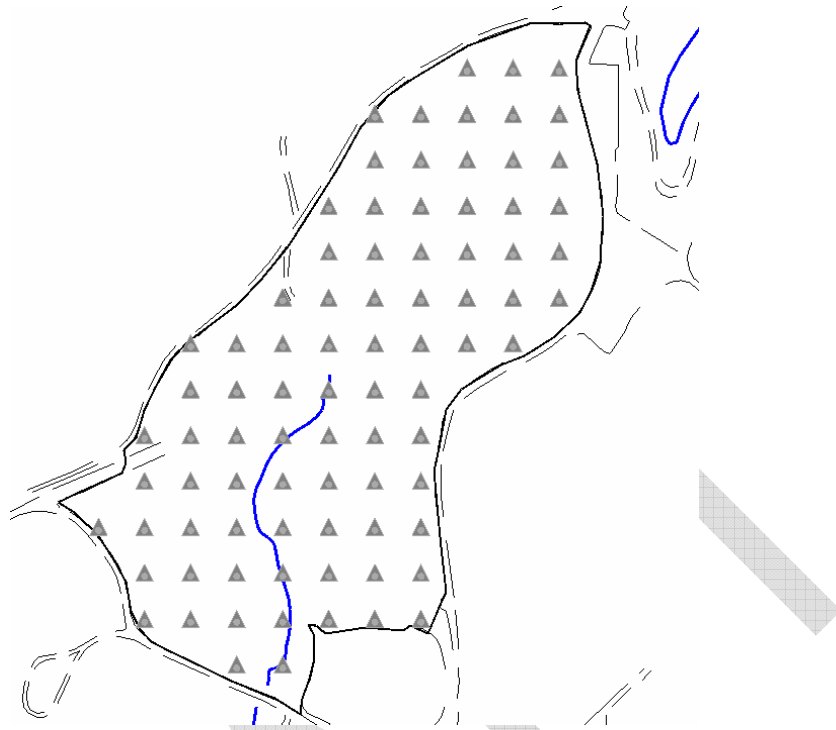
Select a Square grid. The lengths are irrelevant since they will all be the same for a square. In the Shape Definition we will choose a value of 1.0. This is the ratio of the major to minor axis. Therefore, a selection of 1 means that the major/minor = 1 means that the major=minor means that we are talking about a circle. Press the Refresh button to see a preview. Since the object is circular the orientation parameter is irrelevant. Leave it as random. Under the Hot Spot Definition select major radius length and choose 50 (ft). Finally under the probability that a hotspot exists and we miss it select 10%. This means that there is a 90% chance that either it doesn't exist at all or it does and we find it. Press the Show The Results button.

SADA places 37 new samples in accordance with the spatially delineated probability that the elevated area exists at all.



This set of samples taken together produce a 10% that we could be in the situation where the hotspot does exist and we missed it by sampling. Notice in the northern region, there is only a 15% chance the object exists at all. Yet with a 10% limit, we'll need to take at least a few samples. In the southern area, we obviously spend more effort sampling. In fact, the closer the probability of existence is to 1, the closer we are to just a traditional ellipgrid model.

By comparison, the traditional uninformed ellipgrid requires 87 samples to search for a hotspot of the same size with a 90% probability of discovery.



This is a 57% reduction in sampling requirements gained strictly by acknowledging some prior knowledge about the site.