|  |  |
| --- | --- |
| ­­Module Title: **Introduction** |  |
| Agenda Item Number: 1 |  |
| Rationale for Module (WIFM) | R is a free, open source software that is capable of executing graphical and numerical data summaries, statistical and spatial models, and validation and uncertainty assessments. It offers repeatable, statistical methods for soil survey data. |
| Module Objectives | Upon completion of this module, the participant will be able to:   * Describe the structure and history of soil survey data. * Open R GUI, RStudio, and Rcmdr programs. * Import/Export data in R. * Save R files and data. * Install, load, and update R packages. |
| Resources Needed | 1. R and RStudio software (installed prior to class start) 2. Computer and internet access |
| Time Allotted: | \_80-98\_ minutes |
| Bad Weather Plan: | This is an inside activity |

| **Do** | **Say** | **Notes** | **Time/ Minutes** |
| --- | --- | --- | --- |
| * 1. Overview of Module 1 |  | Provide a course overview.  Discuss the objectives of the course.  See slides X - X | 15 |
| * 1. Discuss how R is applicable to soil survey data. | Who has experience using R?  For those who have used R, what do you use it for? | Have students’ type responses in chat.  Reference 1.1 | 3-10 |
| * 1. Overview | Please take a moment to read to yourselves the objectives for this module. | Describe the structure and history of soil survey data.  Open R GUI, RStudio, and Rcmdr programs.  Import/Export data in R.  Save R files and data.  Install, load, and update R packages. | 2-3 |
| * 1. Deliver PowerPoint to introduce R GUI and how it is used in soil survey | Why R? For example: explore spatial data trends (ESI point data, pedon)  Soil scientists are using R for exploring soil data, modeling soil properties or classes, validating and providing uncertainty assessments of raster-based model predictions, and developing and editing packages to expand functionality in R. | See slides X - X  Other examples can be used – state heritage plot data, witness tree data, KSSL, soil survey observations, forest service observations  Reference 1.2 | 10 |
| * + 1. Exercise using Basic R commands | Set working directory after opening R. | This is accomplished individually and includes importing, viewing, and exporting data.  Use the Sand data set.  Reference 1.3 and 1.6 | 10 |
| * 1. Deliver PowerPoint on Data Objects |  | Reference PowerPoint slides X – X  Reference 1.4 | 5 |
| * 1. Installing packages | What are packages? Continuing with the data set from the previous exercise you will install the Rcmdr package, load the sand data set and generate descriptive statistics. | Use Rcmdr as the example.  Reference 1.5 and 1.7 | 10 |
| 1.4.1 Exercise |  | Install all required packages (see script) | 15 |
| * 1. Deliver PowerPoint introducing RStudio |  | See PowerPoints X-X  Reference 1.8 | 10 |
| * 1. Summary of Module 1 | Let’s review what we just went over. | Questions:  What is a package?  A. Collections of code to do specific functions.  What is the first thing you do when you open R.  A. Set working directory  List data objects  A. Vectors, matrices, lists, arrays, and data frames | 5-10 |

|  |  |
| --- | --- |
| Module Title: | The Data We Use |
| Agenda Item Number: | 2 |
| Rationale for Module (WIFM) | Understanding tabular data structure, including data types, accuracy, precision, and assumptions, and its spatial relationship to environmental factors is imperative for data analysis and population. |
| Module Objectives | Upon completion of this module, the participant will be able to:   * Differentiate four data types. * Explain the difference between accuracy and precision. * Create formatted data for use in R. * Establish an ODBC connection to view NASIS data in R. * Determine spatial descriptive statistics in ArcMap, TEUI, and R. |
| Resources Needed | TEUI Add-in required and ArcMap opened prior to the start of module. |
| Time Allotted: | \_81\_ minutes |
| Bad Weather Plan: | This is an inside activity |

| **Do** | **Say** | **Notes** | **Time/ Minutes** |
| --- | --- | --- | --- |
| * 1. Review Module 1. | Let’s review what we covered in Module 1 and what you should be able to do at this point in training. | Describe the structure and history of soil survey data.  Open R GUI, RStudio, Rcmdr programs.  Import/Export data in R.  Save R files and data  Install, load, and update R packages. | 3 |
| * 1. Overview of Module 2. | Please take a moment to read to yourselves the objectives for this module. | Differentiate four data types.  Explain the difference between accuracy and precision.  Create formatted data for use in R.  Establish an ODBC connection to view NASIS data in R.  Determine spatial descriptive statistics in ArcMap, TEUI, and R. | 2 |
| * 1. Identify four data types. | What are the four data types that you typically encounter in your day-to-day work? | * Discuss continuous and discrete data. * Reference 2.1 | 5 |
| * 1. Explain the difference between accuracy and precision. |  | * See PowerPoint X – X * Reference 2.2 | 7 |
| 2.3 Create formatted data for use in R. |  | * Make sure you mention extract multi-values to points * See PowerPoints X - X * Reference 2.3 | 10 |
| 2.3.1 Exercise - Using data from NASIS in R. | Let’s do an exercise. This one is to be done individually. | * Predefined selected set in NASIS completed by all prior to this module. Suggest Coolville, OH (all need to use the same location) | 15 |
| 2.4 Open questions to group. | How do you determine L, RV, and H values for slope gradient? Why? |  | 4 |
| 2.5 Deliver PowerPoint on determining spatial descriptive statistics in ArcMap, TEUI, and R. |  | * See PowerPoints X - X * Reference 2.X | 10 |
| 2.6 Exercise – Zonal Statistics in ArcMap. | It is time for an exercise. | Individual completes exercise on same dataset using zonal statistics in ArcMap. | 15 |
| * 1. Summary |  | Questions:  Vegetation classes are an example of what data type?  A. Nominal  Strong SBK structure is an example of what data type?  A. Ordinal  What data type is slope aspect?  A. Interval  Given two observations with slope aspect of 350 and 10, what is the average slope aspect?  A. 0  The closeness of repeated measurements to each other is called?  A, Precision  Given an example, is it suitable for input? (Show a/b example located XXX slide)  A. Answer will be on following slide  Given what you know now, how will you apply these techniques in your area? | 10 |

|  |  |
| --- | --- |
| Module Title: | Sample Design |
| Agenda Item Number: | 3 |
| Rationale for Module (WIFM) | Sampling is a fundamental part of statistics. The goal is to collect samples that provide an accurate representation of the population under study. |
| Module Objectives | Upon completion of this module, the participant will be able to:   * Explain common sampling schemes and determine the best scheme for applicable situations. * Describe cLHS and use the cLHS tool in TEUI to select a sample set using provided data. * Determine required number of samples for given scenarios. |
| Resources Needed | TEUI Add-in and ArcMap opened prior to the start of module |
| Time Allotted: | \_90\_ minutes |
| Bad Weather Plan: | This is an inside activity |

| **Do** | **Say** | **Notes** | **Time/ Minutes** |
| --- | --- | --- | --- |
| 3.1 Review Module 2. | Let’s review what we covered in Module 2. | Differentiate four data types.  Explain the difference between accuracy and precision.  Create formatted data for use in R.  Establish an ODBC connection to view NASIS data in R.  Determine spatial descriptive statistics in ArcMap, TEUI, and R. | 5 |
| 3.2 Overview of Module 3 | Please take a moment to read to yourselves the objectives for this module. | Explain common sampling schemes and make a determination of the best scheme for applicable situations.  Describe cLHS and use the cLHS tool in TEUI to select a sample set using provided data.  Determine required number of samples. | 2 |
| 3.3 Open questions to group | Question – What sampling schemes have you used?  If you have a need to collect samples, how do you determine what sampling scheme would you use? |  | 3 |
| 3.4 Deliver PowerPoint. Explain common sampling schemes and determine the best scheme for applicable situations. |  | See PowerPoint slides X-X Reference 3.1 thru 3.6 | 10 |
| 3.4.1 Exercise – A horizon thickness, soil depth, surface texture, surface stoniness class, tree productivity, forage productivity | Let’s take 10 minutes to do an exercise in small groups and then we’ll have each group present their answer to the entire class. | Break into small groups with each group having a different scenario. Upon completion, present answer to the class. | 20 |
| 3.5 Deliver PowerPoint. Describe cLHS and use the cLHS tool in TEUI to select a sample set using provided data. |  | See PowerPoint slides X - X  Reference 3.7 – 3.8 | 10 |
| 3.5.1 Exercise – Case Study? | Let’s look at a case study. |  | 15 |
| 3.6 Discuss how we determine required number of samples. |  | Do as a whiteboard activity or as a PowerPoint | 5 |
| 3.6.1 Exercise | Let’s try calculating the required sample size for different scenarios. | After description, participants calculate required sample size in provided scenarios.  t table link needed for scenarios.  Reference 3.9 | 10 |
| 3.7 Summary |  | Questions:  Given a stratified random design, what would be an appropriate data element to assess?  A. Within map unit evaluations, among phases of map units,  List things to consider for each sampling design.  A. Cost, time, accuracy, precision  Compare and contrast considerations between Stratified Random (SR) and Two-Stage Random (2SR) designs.  A. SR is easy to implement compared to 2SR  2SR is more efficient, but may miss the mark  If you knew absolutely nothing about the study area, which design(s) would be appropriate?  A. Simple Random, Systematic, Two-Stage Random  What is a reasonable level of precision for soil survey field work?  A. 70% or more is good. This is the natural, messy world. | 10 |

|  |  |
| --- | --- |
| Module Title: | Exploratory Data Analysis |
| Agenda Item Number: | 4 |
| Rationale for Module (WIFM) | Before embarking on statistical tests and inferences, it is essential to understand your data. This will be done through conventional numerical and graphical methods. John Tukey (Tukey, 1977) advocated the practice of exploratory data analysis (EDA) as a critical part of the scientific process. |
| Module Objectives | Upon completion of this module, the participant will be able to:   * Use basic exploratory data analysis techniques to examine provided data. * Identify trends in data for future applications |
| Resources Needed |  |
| Time Allotted: | \_120\_ minutes |
| Bad Weather Plan: | This is an inside activity |

| **Do** | **Say** | **Notes** | **Time/ Minutes** |
| --- | --- | --- | --- |
| 4.1 Review Module 3 Concepts | In our last module, we went over how to …. | Explain common sampling schemes and determine the best scheme for applicable situations.  Describe cLHS and use the cLHS tool in TEUI to select a sample set using provided data.  Determine required number of samples for given scenarios. | 5 |
| 4.2 Open Module 4 with a question | How do you examine your soil survey data? Is it objective or subjective? | May want to have everyone type answers into the chat and then quickly discuss answers | 5 |
| 4.3 Overview of objectives | Our objectives for this modules are …. | Calculate basic exploratory data analysis.  Identify trends in data for future applications | 5 |
| 4.4 Delivery PowerPoint on calculate basic exploratory data analysis. |  | See PowerPoint slides X - X  Reference 4.0-4.14 | 20 |
| 4.4.1 Exercise – Instructor led | Let’s have everyone do an exercise. | Individuals complete with same scenario for everyone. | 30 |
| 4.5 Deliver PowerPoint on identifying trends in data for future applications |  | See PowerPoint slides X - X  Reference 4.0-4.14 | 15 |
| 4.5.1 Exercise - Instructor led | Let’s try it. | Individuals complete with same scenario for everyone. | 30 |
| 4.6 Summary |  | Questions:  Show a histogram and ask if the distribution is normal.  A. Show a skewed distribution and discuss.  What does inter-quartile range define?  A. Middle 50% of observations  What percentiles approximate 2 SD?  A. 5% & 95%  How would you use a box plot in soil survey work (your area)?  A. Open ended (one categorical, one non-categorical), e.g. land use and A horizon thickness. | 10 |

Start of Advanced Portion of the Course

|  |  |
| --- | --- |
| Module Title: | Clustering and Ordination |
| Agenda Item Number: | 5 |
| Rationale for Module (WIFM) | Provide a means for grouping or separating data. |
| Module Objectives | Upon completion of this module, the participant will be able to:   * Produce and interpret a dendrogram. * Calculate and interpret a distance matrix. * Apply non-metric multidimensional scaling. |
| Resources Needed |  |
| Time Allotted: | \_110\_ minutes |
| Bad Weather Plan: | This is an inside activity |

| **Do** | **Say** | **Notes** | **Time/ Minutes** |
| --- | --- | --- | --- |
| 5.1 Review Module 4. | In our last module, we went over how to …. | Calculate basic exploratory data analysis.  Identify trends in data for future applications | 5 |
| 5.2 Overview of Module 5 | In distance matrix clustering, our objectives are for you to … | Produce and interpret a dendrogram.  Calculate and interpret a distance matrix.  Apply non-metric multidimensional scaling. | 5 |
| 5.3 Open with the question/ask | Name some applications for clustering techniques. |  | 5 |
| 5.4 Deliver PowerPoint on Introduction |  | See PowerPoint X thru X  Reference: | 10 |
| 5.5 Deliver PowerPoint on Produce and interpret a dendrogram. |  | See PowerPoint X thru X  Reference: | 10 |
| 5.5.1 Exercise | Let’s do an exercise in groups. |  | 15 |
| 5.6 Deliver PowerPoint on Calculate and interpret a distance matrix. |  | See PowerPoint X thru X  Reference: | 10 |
| 5.6.1 Exercise | Time for an exercise. |  | 15 |
| 5.7 Deliver PowerPoint on Apply non-metric multidimensional scaling. |  | See PowerPoint X thru X  Reference: | 10 |
| 5.7.1 Exercise | Let’s try it. |  | 15 |
| 5.8 Summary/Review |  | Facilitated discussion of a dendrogram. (See notes and diagram.)  Questions (NEED TO ADD) | 10 |

|  |  |
| --- | --- |
| Module Title: | Regression |
| Agenda Item Number: | 6 |
| Rationale for Module (WIFM) | Reinforce concept of dependent and independent variables and their application to soil survey in a linear framework. |
| Module Objectives | Upon completion of this module, the participant will be able to:   * Compute and interpret coefficients in a linear regression analysis in R. * Interpolate regression model in R to produce a raster layer. |
| Resources Needed |  |
| Time Allotted: | \_85\_ minutes |
| Bad Weather Plan: | This is an inside activity |

| **Do** | **Say** | **Notes** | **Time/ Minutes** |
| --- | --- | --- | --- |
| 6.1 Review 1st half of course | Let’s review what we covered in the online portion of this course ….  Are there any questions | See objectives for first 5 modules | 15 |
| 6.2 Overview of Module 6 | Now we are going to talk about regression. Our objectives for this module are: | Compute and interpret coefficients in a linear regression analysis in R.  Interpolate regression model in R to produce a raster layer. | 5 |
| 6.3 Open with a question or ask … | Describe an example of how regression has been used in the realm of natural resources |  | 5 |
| 6.4 Deliver PowerPoint on computing and interpreting coefficients in a simple linear regression and multiple linear regression analysis in R. |  | See PowerPoint  Mention pedotransfer functions as an example  Reference: 6.1 thru 6.4 | 20 |
| 6.4.1 Exercise – Linear Regression | In this exercise, I want you to:   * Create a multiple linear regression model * Report descriptive statistics * Interpolate model to raster layer | Data provided  Individual exercise | 30 |
| 6.5 Summary/Review |  | Questions:  What are the components of simple linear regression equation?  A. 1 independent and 1 dependent variable, slope, and Y intercept  What is the main limitation of using linear regression?  A. Normality  Can you think of other names for regression?  A. Pedotransfer function, stepwise analysis, linear models  Are more variables better?  A. (generates discussion) not necessarily  \* Note – Talk about R square inflation and consider ockham’s razor. | 10 |

|  |  |
| --- | --- |
| Module Title: | Logistic Regression |
| Agenda Item Number: | 7 |
| Rationale for Module (WIFM) | Provides a means of predicting binomial phenomenon. |
| Module Objectives | Upon completion of this module, the participant will be able to:  - Compute and interpret a logistic regression model in R.  - Interpolate logistic regression model in R to produce a raster layer. |
| Resources Needed | Have the computer lab set up with software prior to student arrival. |
| Time Allotted: | \_75\_ minutes |
| Bad Weather Plan: | This is an inside activity |

| **Do** | **Say** | **Notes** | **Time/ Minutes** |
| --- | --- | --- | --- |
| 7.1 Overview of module 7 | In our logistic regression module, our objectives are to have you …. | Compute and interpret a logistic regression model in R.  Interpolate logistic regression model in R to produce a raster layer. | 5 |
| 7.2 Open with this question | What is a binomial response or data type that is used in your work? | Examples for question:  Soil feature presence/absence  Target soil class/non-target   classes | 5 |
| 7.3 Deliver PowerPoint on compute and interpret a logistic regression model in R. |  | See PowerPoint X thru X  Reference 7.1 thru 7.3 | 15 |
| 7.3.1 Exercise | In this exercise, I want you to individually:  Create a logistic regression model  Report descriptive statistics  Interpolate model to raster layer | Use common data set  Perform exercise individually  Set output parameter to be a probability surface. | 40 |
| 7.4 Summary/Review |  | Questions:  Can you think of a way to use the output in your work?  Name some advantages of this technique.  A. Works with non-normal data, binomial.  Contrast with linear regression  A. Works with non-normal data, binomial, output raster is a probability surface. | 10 |

|  |  |
| --- | --- |
| Module Title: | Tree-based Models |
| Agenda Item Number: | 8 |
| Rationale for Module (WIFM) | Tree-based models are a common technique for digital soil mapping. Many soil survey problems lend themselves to tree-based techniques. |
| Module Objectives | Upon completion of this module, the participant will be able to:   * Create and evaluate CART and Random Forest models in R. * Interpolate tree-based models in R to a raster output. |
| Resources Needed |  |
| Time Allotted: | \_90\_ minutes |
| Bad Weather Plan: | This is an inside activity |

| **Do** | **Say** | **Notes** | **Time/ Minutes** |
| --- | --- | --- | --- |
| 8.1 Review Module 7 |  | Compute and interpret a logistic regression model in R.  Interpolate logistic regression model in R to produce a raster layer. | 5 |
| 8.2 Overview of Module 8 | Everyone take a moment to read the objectives for this module. | Create and evaluate CART and Random Forest models in R.  Interpolate tree-based models in R to a raster output. | 2 |
| 8.3 Open with the question | How have you, or could you, use decision trees? |  | 3 |
| 8.4 Deliver PowerPoint on creating and evaluating CART models in R. |  | See PowerPoint X thru X  Reference 8.1 - 8.3 | 10 |
| 8.4.1 Exercise | In this exercise, I want you to:   * Create a classification and a regression tree model using rpart * Report summary statistics | Use common data set  Perform exercise individually  Mention output parameters (class vs. prob.) | 20 |
| 8.5 Deliver PowerPoint on creating and evaluating random forest models in R. |  | See PowerPoint X thru X  Reference 8.5 – 8.7 | 10 |
| 8.5.1 Exercise | In this exercise, I want you to:   * Create a random forest regression model * Report summary statistics | Use common data set  Perform exercise individually | 10 |
| 8.5.2 Exercise | In this exercise, I want you to:  - Interpolate a rpart and a randomForest model to a raster layer | Use models created in the last two exercises  Perform exercise individually | 20 |
| 8.6 Summary/Review |  | Questions:  Why is random forest often referred to as double random?  A. Number of covariates and samples  How does random forest perform internal validation?  A. Out of bag samples  True or false – numerical and categorical data can be used in the creation of a classification tree?  A. True  Contrast CART with random forest.  A. the tree you can see versus the one you can’t. | 10 |

|  |  |
| --- | --- |
| Module Title: | Case Study |
| Agenda Item Number: | 9 (Encompasses Modules 4-8) |
| Rationale for Module (WIFM) | Reinforces material previously covered and provides opportunity to apply information in real world scenario. |
| Module Objectives | Upon completion of this module, the participant will be able to:   * Given a scenario, evaluate dataset, select and apply a model(s), justify decisions, and explain results. |
| Resources Needed |  |
| Time Allotted: | \_95\_ minutes |
| Bad Weather Plan: | This is an inside activity |

| **Do** | **Say** | **Notes** | **Time/ Minutes** |
| --- | --- | --- | --- |
| 9.1 Review via case study in groups | We are going to do a short review on Modules 4-8 and then working in groups do a case study on what we have covered. | - Do a peer review with class over modules 4-8. (15 minutes to develop and 15 minutes to present)  - Have participants perform EDA techniques in groups. (15 min)  - Discuss EDA findings with entire class. (20 min)  - Have participants model a scenario with groups’ chosen method (20 min)  - Discuss model findings with entire class. – Pick group(s) to discuss. (10 min) | 95 |

|  |  |
| --- | --- |
| Module Title: | Validation and Uncertainty |
| Agenda Item Number: | 10 |
| Rationale for Module (WIFM) | Provide the tools to allow proper evaluation of models to guide decision making. |
| Module Objectives | Upon completion of this module, the participant will be able to:   * Differentiate between internal and external validation techniques. * Describe sources of uncertainty. * Quantify model uncertainty. * Interpret validation statistics. |
| Resources Needed |  |
| Time Allotted: | \_90\_ minutes |
| Bad Weather Plan: | This is an inside activity |

| **Do** | **Say** | **Notes** | **Time/ Minutes** |
| --- | --- | --- | --- |
| 10.1 Review previous material |  |  | 5 |
| 10.2 Overview of Validation and Uncertainty | Everyone take a moment to read to yourselves the objectives for this module… | Differentiate between internal and external validation techniques.  Describe sources of uncertainty.  Quantify model uncertainty.  Interpret validation statistics. | 5 |
| 10.3 Open with the questions | What does uncertainty mean?  What does validation mean? | Create and calculate examples using pencil and paper, Excel and R. | 5 |
| 10.4 Deliver PowerPoint |  | See PowerPoint X thru X  Reference: still working on this section. | 10 |
| 10.5 Deliver PowerPoint on Sources of Error |  | See PowerPoint X thru X  Discuss examples used in previous modules.  Reference: 9.1 | 10 |
| 10.6 Deliver PowerPoint on Error and Uncertainty |  | See PowerPoint X thru X  Discuss examples used in previous modules.  Reference: 9.2 | 10 |
| 10.7 Deliver PowerPoint on Validation Methods |  | See PowerPoint X thru X  Discuss examples used in previous modules.  Reference: 9.3 | 10 |
| 10.7.1 Exercise | Now let’s do an exercise in groups. | Creation and interpretation of a confusion matrix will be done by each group. Validation and uncertainty will be calculated and assessed. | 20 |
| 10.8 Summary/Review |  | Questions:  All models are wrong? True/False  A. True  What’s wrong with always using r2 or p-values to determine model accuracy?  A. r2 inflation, not a controlled experiment, normality, number of samples.  Measuring soil depth, with multiple transects:  A linear regression model was used to predict soil depth. How would you validate the model?  What validation technique would you use and why?  r2 was 0.85 and RMSE was 0.6. Uncertainty revealed that ridge tops showed higher errors. There is no time or money left for further study. What advice do you provide the Conservationist? | 15 |