NEON Algorithm Theoretical Basis Document: TOS Small Mammal Abundance and Diversity - QA/QC of Raw Field and Lab Data

PREPARED BY	ORGANIZATION	DATE
Katherine M. Thibault	FSU	07/21/2014

APPROVALS (Name)	ORGANIZATION	APPROVAL DATE

RELEASED BY (Name)	ORGANIZATION	RELEASE DATE

See configuration management system for approval history.

Change Record

REVISION	DATE	ECO #	DESCRIPTION OF CHANGE
A_DRAFT	05/14/2014	ECO-01880	Initial Draft Release

TABLE OF CONTENTS

LIST OF TABLES AND FIGURES

1 DESCRIPTION

1.1 Purpose

This document details the algorithms used for creating a subset of NEON Level 1 data products that are the quality controlled products generated from raw Level 0 data, and associated metadata. In the NEON data products framework, the raw data collected in the field, for example, hindfoot length of an individual small mammal, are considered the lowest level (Level 0). Raw data that have been quality checked via the algorithms detailed herein, as well as simple metrics that emerge from the raw data, such as total species richness of small mammals at a particular site, are considered Level 1 data products. This document relates only to the former group of L1 data products, the quality controlled pass-through products from the Level 0 data products.

This document includes a detailed discussion of measurement theory and implementation, appropriate theoretical background, data product provenance, quality assurance and control methods used, approximations and/or assumptions made, and a detailed exposition of uncertainty resulting in a cumulative reported uncertainty for this product.

1.2 Scope

This document describes the theoretical background and entire algorithmic process for creating a subset of quality controlled and calibrated L1 data products from input data. These data products comprise the small mammal box trapping data (NEON.DOM.SIT.DP1.10072). This document does not provide computational implementation details, except in cases where these stem directly from algorithmic choices explained here. This document also provides details relevant to the publication of the data products via the NEON data portal (NEON Data Publication Workbook for TOS Small Mammal Abundance and Diversity: QA/QC of Raw Field and Lab Data (AD[12])).

This document describes the algorithms for ingesting and performing automated quality assurance and control procedures on field-collected data pertaining to small mammal abundance and diversity, and, by extension, mammal-borne pathogens (see AD[10] for ingest ATBD for pathogen analytical data). The raw data that are processed in this document are detailed in the NEON Raw Data Ingest Workbook for TOS Small Mammal Abundance and Diversity (AD[11]).

2 RELATED DOCUMENTS AND ACRONYMS

2.1 Applicable Documents

AD[01]	NEON.DOC.000001	NEON Observatory Design (NOD) Requirements
AD[02]	NEON.DOC.005003	NEON Scientific Data Products Catalog
AD[03]	NEON.DOC.005004	NEON Level 1-3 Data Products Catalog
AD[04]	NEON.DOC.005005	NEON Level 0 Data Product Catalog
AD[05]	NEON.DOC.005011	NEON Coordinate Systems Specification
AD[06]	NEON.DOC. 001247	NEON ATBD: QA/QC Data Validation and Plausibility Testing of TOS and AOS Field and Lab Data
AD[07]	NEON.DOC.002162	NEON Algorithm Theoretical Basis Document: Taxonomic Consensus Identifications and Uncertainty
AD[08]	NEON.DOC.000915	TOS Science Design for Small Mammal Abundance and Diversity
AD[09]	NEON.DOC.000481	TOS Protocol and Procedure: Small Mammal Sampling
AD[10]	NEON.DOC.001240	NEON Algorithm Theoretical Basis Document: Rodent-borne Pathogens - QA/QC of Raw Field and Lab Data and Prevalence Measure Calculations
AD[11]	NEON.DOC.001406	NEON Raw Data Ingest Workbook for TOS Small Mammal Abundance and Diversity
AD[12]	NEON.DOC.001417	NEON Data Publication Workbook for TOS Small Mammal Abundance and Diversity: QA/QC of Raw Field and Lab Data
AD[13]	NEON.DOC.004309	NEON Field Site Information
AD[14]	NEON.DOC.002261	TOS Spatial Data
AD[15]	NEON.DOC.002259	NEON Taxonomic Name List for Small Mammals
AD[16]	NEON.DOC.002253	List of Valid Identification Qualifiers

2.2 Reference Documents

RD[01]	NEON.DOC.000008	NEON Acronym List
RD[02]	NEON.DOC.000243	NEON Glossary of Terms
RD[03]	NEON.DOC.000911	NEON TOS Science Design for Vectors and Pathogens

3 DATA PRODUCT DESCRIPTION

Small mammal abundance and diversity will be sampled at regular intervals by NEON field technicians at core and relocatable sites. Sampling is based on the lunar calendar, with timing of sampling constrained to occur within 10 days before or after the new moon. Sampling may occur as frequently as every new moon (i.e., 12 or 13 times per year) to as infrequently as 4 times per year (i.e., every other new moon from April – October), depending on logistical constraints (e.g., deep snow, limited labor resources). Small mammals are sampled using box traps (models LFA, XLK, H.B. Sherman Traps, Inc., Tallahassee, FL, USA) arrayed in three – eight 10 x 10 grids (Figure 1), depending on the size of the site. For additional details on the sampling design and associated protocol, see the TOS Science Design for Small Mammal Abundance and Diversity (AD[08]) and the TOS Protocol and Procedure: Small Mammal Sampling (AD[09]). Products resulting from this sampling include the species identification and unique identifier for each individual captured (excluding captures of non-target species; AD[09]), as well as a suite of standard size measurements and reproductive condition data.

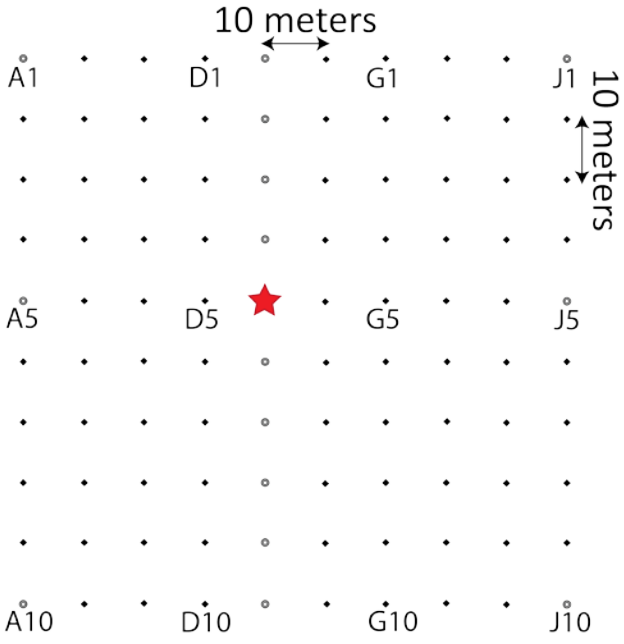

Figure 1. Design of the small mammal trapping grid, consisting of 100 box traps separated by 10 m. Trap coordinates are designated by a unique letter and number combination (e.g., A1, G10). The plot-level coordinate is trap E5, designated by the red star.

3.1 Variables Reported

This ATBD describes the steps needed to generate the L1 data product, small mammal box trapping (NEON.DOM.SIT.DP1.10072). The small mammal box trapping data product is the mark-recapture, demographic, and size data collected by box trapping for small mammals (i.e., rodents (Rodentia) and shrews (Soricomorpha) < 600 grams).

Subproducts of this data product are listed below (Table 1). Detailed lists of the associated subproducts and metadata products are provided separately, along with example data in publication-ready spreadsheets (NEON Data Publication Workbook for TOS Small Mammal Abundance and Diversity: QA/QC of Raw Field and Lab Data (AD[12])). Field names have been standardized with Darwin Core terms (http://rs.tdwg.org/dwc/; accessed 23 August 2013), the Global Biodiversity Information Facility vocabularies (http://rs.gbif.org/vocabulary/gbif/; accessed 27 August 2013), and with the Bird Monitoring Data Exchange standards (http://www.avianknowledge.net; accessed 23 August 2013), whenever possible. Geospatial data shall conform to the standards set forth in the NEON Coordinate Systems Specification (AD[05]).

Table 1. List of subproducts produced in this ATBD in the data product, small mammal box trapping (NEON.DOM.SIT.DP1.10072). The list is not exhaustive and a variety of supporting data will also be made available.

Number	Field Name	Description
NEON.DOM.SITE.DP1.10072.001.00047.003.001.001	individualID	Domain-level unique identifier based on domain number: NEON.MOD.DXX.123456
NEON.DOM.SITE.DP1.10072.001.00012.003.001.001	taxonID	4-character species code
NEON.DOM.SITE.DP1.10072.001.00055.003.001.001	scientificName	Scientific name, associated with the taxonID. This is the name of the lowest level taxonomic rank that can be determined.
NEON.DOM.SITE.DP1.10072.001.00056.003.001.001	taxonRank	The taxonomic rank of the most specific name in the scientificName.
NEON.DOM.SITE.DP1.10072.001.00013.003.001.001	identificationQualifier	A standard term to express the determiner's doubts about the Identification.
NEON.DOM.SITE.DP1.10072.001.00057.003.001.001	identificationStatus	A categorical indicator of changes applied to a taxonomic identification based on conflicting sources, where applicable.
NEON.DOM.SITE.DP1.10072.001.00014.003.001.001	sex	Sex of the individual; M for male, F for female
NEON.DOM.SITE.DP1.10072.001.00058.003.001.001	reproductiveCondition	The reproductive condition of the individual at the time of capture. R for reproductive; N for non-reproductive
NEON.DOM.SITE.DP1.10072.001.	recapture	Indicates whether or not the captured

00027.003.001.001		individual is a recapture; 'Y' for yes, 'N' for no
NEON.DOM.SITE.DP1.10072.001.00028.003.001.001	fate	The fate of the individual, unless marked and released; dead = dead, escaped = escaped while handling, nontarget = released, non-target species
NEON.DOM.SITE.DP1.10072.001.00015.003.001.001	lifeStage	The age class of the individual at the time the Occurrence was recorded. juvenile = obvious signs of a very young individual, small size, distinctive pelage coloration; subabult; adult
NEON.DOM.SITE.DP1.10072.001.00016.003.001.001	testes	Condition of the testes at time of capture; if mature: scrotal = testes descended, nonscrotal = testes abdominal
NEON.DOM.SITE.DP1.10072.001.00017.003.001.001	nipples	Condition of the nipples at time of capture; if mature: enlarged = nipples enlarged, nonenlarged = nipples not enlarged
NEON.DOM.SITE.DP1.10072.001.00018.003.001.001	pregnancyStatus	Condition at time of capture; if mature: pregnant
NEON.DOM.SITE.DP1.10072.001.00019.003.001.001	vagina	Condition of the vagina at time of capture; if mature: swollen, plugged, neither
NEON.DOM.SITE.DP1.10072.001.00020.003.001.001	hindfootLength	length of left hindfoot; including claws; in millimeters
NEON.DOM.SITE.DP1.10072.001.00021.003.001.001	earLength	length of left ear; in millimeters
NEON.DOM.SITE.DP1.10072.001.00022.003.001.001	tailLength	length of tail; in millimeters
NEON.DOM.SITE.DP1.10072.001.00023.003.001.001	totalLength	total length (head + body); in millimeters
NEON.DOM.SITE.DP1.10072.001.00024.003.001.001	weight	Live weight; as measured with a spring scale; in grams
NEON.DOM.SITE.DP1.10072.001.00043.001.001.001	samplingEffort	The total amount of trapping effort expended during a sampling bout (eventID); with corrections made for disturbed traps.

3.2 Temporal Resolution and Extent

The finest temporal resolution at which small mammal data will be tracked is night of trapping (i.e., date – YYYY-MM-DD, and night within bout – 1, 2, 3). Three consecutive (± 5 days) nights of trapping at a site will then be aggregated to comprise a sampling bout (eventID), the temporal unit that will be used for

calculating density estimates and other higher level data products. The total number of bouts per year is planned to vary with site designation (core vs. relocatable) and site-specific duration and severity of weather conditions (e.g., trapping will not occur during bouts of extreme cold, high snow pack, or the combination of cold temperatures and precipitation). It is expected that 3 - 12 bouts of trapping will occur annually, depending on the site.

3.3 Spatial Resolution and Extent

The finest spatial resolution at which small mammal data will be tracked is trap location (i.e., trapCoordinate). Although summary statistics will not be produced at this scale, the geographic distances traveled by an individual between recapture events will be calculated as part of the algorithm for spatially explicit capture-recapture density estimates (to be detailed in a separate ATBD, in preparation). One trapping plot consists of 100 trap locations (see Figure 1 above); the geographic coordinates for these locations are produced using a Geographic Information System prior to the initiation of sampling. NEON field technicians use a combination of high-resolution GPS units, recreational GPS units, and measuring tape to demarcate the sampling grids, with an uncertainty associated with each trap location of approximately 2 meters. Technicians then place traps within a 1-m radius of these locations during each sampling bout; this spatial variation is not captured. Capture data will be aggregated at the plot scale; i.e., the plot will be treated as a sampling replicate for community- and population-level data products. The total number of plots per site varies with area of the site and the site's habitat heterogeneity, from three to eight plots per site. Data from these plots will be aggregated to yield site-level estimates of community and population parameters for small mammals. Overall, this results in a spatial hierarchy of:

trapCoordinate → plotID → habitat (NLCD veg type) → siteID → domainID

3.4 Associated Data Streams

All of the above data are also directly linked to the rodent-borne pathogen data, as described in the NEON Algorithm Theoretical Basis Document: Rodent-borne Pathogens - QA/QC of Raw Field and Lab Data and Prevalence Measure Calculations (AD[11]).

3.5 Product Instances

The raw field data generated from box trapping for small mammal abundance and diversity are predominantly per individual per night of trapping (per bout, per plot, per site). Small mammal traps are planned to be set at all core and relocatable sites throughout NEON every year. Sampling effort for small

mammal trapping is typically measured in units of trap nights (e.g., 10 trap nights = one trap set for 10 nights or 10 traps set for one night). Total sampling effort, small mammal densities, and small mammal capture rates will vary across NEON sites, and, consequently, the number of individual records of these products to be generated is unknowable.

4 SCIENTIFIC CONTEXT

4.1 Theory of Measurement/Observation

Small mammal field studies have played a key role throughout the history and development of the field of ecology, particularly in the subdisciplines of behavioral, population, and community ecology (Stapp 2010). This is, in part, because small mammals are abundant in virtually all ecosystems, from harsh deserts to arctic and alpine tundra (Merritt 2010). Moreover, box trapping of nocturnal small mammals allows for the simultaneous study of species-specific demography and population sizes, species-specific prevalence of rodent-borne diseases important to public health, and community composition and species diversity. In support of NEON's mission, these measures can be ultimately linked to land use and climate changes, and therefore provide useful metrics of responses in biodiversity to these and other drivers (Kao et al. 2012). NEON small mammal sampling will assess the abundance and diversity of the nocturnal small mammal communities of North America, including the population dynamics of species that are competent reservoirs for infectious disease, at up to 60 sites (depending on study design and associated logistics) throughout North America, Hawaii, and Puerto Rico for a period of 30 years. NEON will use mark-recapture methods to assess the community, population, and disease dynamics of small mammals across time and space (Ostfeld and Parmenter 2008). For more information regarding the scientific basis of the NEON small mammal sampling design, please refer to the NEON TOS Science Design for Small Mammal Abundance and Diversity [AD[08]].

The data products and metadata referenced herein represent standard measurements associated with live-trapping of small mammals for population and community studies (Wilson et al. 1996). For example, hindfoot, ear, tail and total length aid in the determination of species identifications, which are important for the evaluation of community diversity and structure. In addition, weight can be used to approximate the energetic demands of each individual, using established allometric relationships (e.g., Nagy et al. 1999, Thibault et al. 2010), to examine the potential role of resource use in observed population and community dynamics. Furthermore, population parameters, such as the numbers of individuals of various ages and percentage of reproductive adults in a population, are useful for modeling species' populations through time. Finally, mark-recapture data are critical to computing robust density estimates (White and Burnham 1999, Efford et al. 2009, Royle et al. 2009), as well as assessing additional population parameters, such as survivorship and recruitment (e.g., Frase et al. 1990, Reed and Slade 2007).

4.2 Theory of Algorithm

This document describes the algorithms for assessing the integrity of the L0 data stream generated by the field sampling of small mammal abundance and diversity. The approaches described herein are simple yet necessary components of quality control and quality assurance, and include such processing steps as verifying that all required data are recorded for each capture event and tracking the individual-level data for consistency and accuracy through time. The validation steps that are implemented as the data are entered, such as constraining data values to pre-defined ranges, are detailed in the Appendix.

4.2.1 Summary of Algorithm for the Capture Data

1. Check for duplicate data.
2. Separate trap status and capture data into two tables for publication.
3. Assign scientificName and taxonRank values using the reported taxonID and the taxonomic name list for small mammals lookup table (AD[15]).
4. Assign a value for reproductiveCondition based on entries into the testes or the nipples, pregnancyStatus, and vagina fields.
5. Generate unique values for individualID of the form: NEON.MOD.DXX.123456, where MOD = MAM, DXX = domain number, and 123456 represents an auto-increment number that is unique to each individual.
6. Check for data consistency on recaptured individuals: verify that the same individualID corresponds to the same taxonID and sex through time, and that entries for lifeStage are consistent with expected changes through time for each individual.

4.2.2 Summary of Algorithm for Trapping Effort Data

1. Assign eventID and night values to each date – plotID combination:
    a. Generate eventID of the form: MAM.SITE.year.bout number
    b. Night values should be set to 1, 2, or 3, indicating the first night of trapping on a specified plot within a bout, the second night, and the third night, respectively.
2. Calculate setTraps = count of traps that were set on the previously evening.
3. Calculate disturbedTraps = count of traps that were disturbed, visited but not triggered, or closed without a capture per plotID per date.
4. Calculate samplingEffort = (sum of trapsSet for all sampling dates within an eventID at a site) – (sum of disturbedTraps for all sampling dates within an eventID at a site)
5. Check for duplicate records.
6. Check that the sampling bout was sufficiently documented, with three dates per plot per site having at least one record.

4.3 Special Considerations

The small mammal data are unique among the TOS animal data products, in that (1) individuals are captured, marked, and then released, producing repeated measures from the same individuals, and (2) the vast majority of taxonomic identifications are done by parataxonomists, i.e., domain staff conducting the live trapping in the field. However, a subset of tissues collected from these released individuals will be sequenced to employ a DNA barcoding approach to verify species-level identifications (e.g., Janzen and Hallwachs 2011, Gibson et al. 2012). Moreover, as a complement to the mark-recapture data, voucher specimens of all resident species will be collected from each domain, with a target minimum of 5 vouchers per species collected every 5 years. Incidental losses due to trapping and handling will all be vouchered as well. These vouchers will be archived at accredited institutions; as such, their taxonomic identification will be reassessed by an expert at the institution. The algorithm for generating and synthesizing these data streams into a consensus taxonomic identification for each record, with associated uncertainty, will be added in subsequent versions of this document, with functions detailed in a separate document (AD[07]).

5 ALGORITHM IMPLEMENTATION

Throughout the algorithm implementation section of this ATBD, 'nodata', 'null', and/or 'NA' indicates a blank cell. All variables reported from the field or laboratory technician (L0 data) are listed in the data ingest workbook (AD[11]), notated here as *mam_perplotnight_in, mam_pertrapnight_in*. Unless otherwise specified in the algorithm below, all variables that appear in tables *mam_perplotnight_pub, mam_capturedata_pub,* and *mam_pertrapnight_pub* (L1 data) have been passed directly from the L0 variables with the same name, as listed in the data publication workbook (AD[12]).'

5.1 Automated Processing Steps for Validation of Field Collected Data

5.1.1 All Data: Generate eventID

1. Determine **bout number** associated with a given sampling date
    a. Use the mean moon phase duration of 29.5305882 to determine the bout number for a given sampling date, where the bout number is the number of the closest new moon (from 01 to 13) within the calendar year.
    b. Sampling is designed to occur within 10 days (before or after) of the new moon, but any date that occurs within 14 days (before and after) of the new moon shall be assigned a bout number.
2. For each row in *mam_perplotnight_in* and *mam_pertrapnight_in*:
    a. Concatenate: 'MAM.' + **siteID** + '.' + year (first four characters in **date** field) of + '.' + **bout number**

b. Insert string into **eventID** field of *mam_perplotnight_pub*, *mam_capturedata_pub,* and *mam_pertrapnight_pub*
c. Example: MAM.CPER.2013.08

5.1.2 All Data: Check for duplicate records

1. For mam_perplotnight_in:
    a. List of fieldNames = [(**plotID, date**)]
    b. Create quality flag field and populate with zeroes
        i. QF Name = [duplicateNightQF]
    c. Check for duplicate records, based on exact matches in the fields listed in (a). Use only the date, even if time is also provided.
    d. If duplicate records are identified:
        i. If values in all remaining fields (e.g., **samplingProtocol, identificationReferences**) *except* uid are also the same:
            A. Pass only one of the records into mam_perplotnight_pub
            B. Enter a '1' into the corresponding **duplicateNightQF**
        ii. Else:
            A. Concatenate conflicting values into the corresponding L1 field
            B. Enter a '1' into the corresponding **duplicateNightQF**.
2. For mam_pertrapnight_in:
    a. List of fieldNames = [(**plotID, date, trapCoordinate**)]
    b. Create quality flag field and populate with zeroes
        i. QF Name = [duplicateTrapCoordinateQF]
    c. Check for duplicate records, based on exact matches in the fields listed in (a). Use only the date, even if time is also provided.
    d. If duplicate records are identified:
        i. If tagIDs are not Null AND tagIDs match across records:
            A. If values in all other fields of the records are identical, then pass only 1 record to the L1 and set its **duplicateTrapCoordinateQF** to '1'.
            B. If there are conflicting values in any of the other fields of these records, pass all records to L1 and set **duplicateTrapCoordinateQF** in all records to '2'.
        ii. Else, if tagIDs are Null OR tagIDs do NOT match across records:
            A. If mam_pertrapnight_in.**trapStatus** == 4 for all records:
                I. Pass all records into mam_capturedata_pub
                II. Enter a '0' into the corresponding **duplicateTrapCoordinateQF** in mam_pertrapnight_pub
            B. Else:

I. Pass all records into mam_pertrapnight_pub and mam_capturedata_pub (if applicable).
II. Enter a '2' into the corresponding **duplicateTrapCoordinateQF** of all duplicate records.

5.1.3 Per Plot Night Data: Run the following processing steps for data in *mam_perplotnight_in*

1. Assign **night** values to each record:
    a. For each unique value of **plotID**:
        i. Determine the unique values of **date** for each **eventID** and order consecutively
        ii. Assign a value of **night** in [*mam_perplotnight_in*] according to the order of the dates: 1, 2, or 3, indicating the first night of trapping on a specified plot within a bout, the second night, and the third night, respectively.
2. Generate **setTraps**
    a. For each unique **date.plotID** combination in *mam_pertrapnight_in*:
        i. Count the number of records in which **trapStatus** >1
        ii. Insert this count into the **setTraps** field of *mam_perplotnight_pub*
3. Generate **disturbedTraps**
    a. For each unique **date.plotID** combination in *mam_pertrapnight_in*:
        i. Count the number of records in which **trapStatus** field contains 2 or 3
        ii. Insert this count into the **disturbedTraps** field of *mam_perplotnight_pub*
4. Generate **samplingEffort**
    a. For each unique **date.plotID** combination in *mam_perplotnight_pub:*
        i. **samplingEffort** = **setTraps** - **disturbedTraps**
5. Check that the sampling bout was sufficiently documented:
    a. Create quality flag field and populate with zeroes
        i. QF Name = [missingRecordsPerBoutQF]
    b. For each **eventID,** count the number of unique **date.plotID** combinations in *mam_perplotnight_in.* If value occurs < 3 times:
        i. insert '1' into **missingRecordsPerBoutQF** of all existing records for the given eventID

5.1.4 Per Trap Night Data: Generate *mam_pertrapnight_pub* and *mam_capturedata_pub*

1. Subset mam_pertrapnight_in to the following fields, and insert all records into mam_pertrapnight_pub:
    a. uid
    b. domainID
    c. siteID
    d. plotID
    e. date (using only the date and not the time, if available)

f. trapCoordinate
g. trapStatus
h. Note that one field of this table, duplicateTrapCoordinateQF, was generated in a previous step

2. Subset mam_pertrapnight_in to all records where **trapStatus** == 4 or 5, including all fields *except* nightUID, into mam_capturedata_pub

5.1.5 Capture Data: Run the following processing steps for data in *mam_capturedata_pub*

1. Remove the time, if present, from the **date** field
2. Generate **individualID**
    a. Generate unique values for **individualID** of the form: NEON.MOD.DXX.123456, where MOD = 'MAM', DXX = domain number, and 123456 represents an auto-increment number that is unique to each individual, based on unique combinations of {**domainID, tagID**} where **tagID** is not Null.
    b. If **tagID** is Null, **individualID** should be Null.
3. Generate **reproductiveCondition** value for each record
    a. Generate **reproductiveCondition** field in *mam_capturedata_pub*
    b. For each row in *mam_capturedata_pub*:
        i. If values in **testes** and **nipples** and **pregnant** and **vagina** fields of *mam_capturedata_pub* are neither NULL nor 'nonscrotal', 'nonenlarged', 'neither':
            A. Insert 'R' into cell in **reproductiveCondition** field of *mam_capturedata_pub*
        ii. Else:
            A. Insert 'N' into cell in **reproductiveCondition** field of *mam_capturedata_pub*
4. Generate **identificationStatus** field and leave blank (the algorithm to populate this field is still under development (AD[07]))
5. Assign L1 **scientificName** and **taxonRank** values
    a. Insert the corresponding values of scientificName and taxonRank using the reported taxonID and the taxonomic name list for small mammals lookup table (AD[15]).
6. Assign spatial metadata values
    a. Insert corresponding values of **nlcdClass, decimalLatitude, decimalLongitude, elevation, coordinateUncertainty, elevationUncertainty** values for each **plotID** and **trapCoordinate** combination (where **subType =** 'mammal'), using the TOS plot_data lookup table (AD[14]).
7. Check for internal and temporal consistencies in the **individualID** and related fields, testing only data collected from the same site for up to the last 5 years:
    a. Create quality flag fields and populate with zeroes

i. QF Names = [**consistencyTagIDTaxonIDQF**; **consistencyTagIDLifeStageQF**; **consistencyTagIDSexQF**; **orderTagIDRecaptureQF**; **orderTagIDLifeStageQF**]

b. If **individualID** Is Null:
   i. Insert -1 into corresponding cell in **consistencyTagIDTaxonIDQF** field
   ii. Insert -1 into corresponding cell in **consistencyTagIDLifeStageQF** field
   iii. Insert -1 into corresponding cell in **consistencyTagIDSexQF** field
   iv. Insert -1 into corresponding cell in **orderTagIDRecaptureQF** field
   v. Insert -1 into corresponding cell in **orderTagIDLifeStageQF** field

c. Else, if the **individualID** is not found in any previous records:
   i. If **recapture** = 'N' (not a recapture), insert '0' into **orderTagIDRecaptureQF** of the current record.
   ii. Else, if **recapture** is Null, assign a value of 'N' to the **recapture** field and insert '0' into **orderTagIDRecaptureQF** of the current record.
   iii. Else, if **recapture** = 'Y':
      A. Insert '1' into **orderTagIDRecaptureQF**.

d. Else, if the individualID is found in a previous record:
   i. If **recapture** = 'N' (not a recapture):
      A. If taxonID is the same in the current record as in the previous record(s):
      B. Change **recapture** to 'Y' and insert '0' into **orderTagIDRecaptureQF** of the current record
   ii. Else, insert '1' into **orderTagIDRecaptureQF** of the current record

f. Then:
   i. Check that **taxonID** and **sex** designations for each individualID do not change from one capture instance to the next.
      A. If **taxonID** is Null:
         I. Insert -1 into corresponding cell in **consistencyTagIDTaxonIDQF** field
         II. Else, if **taxonID** changes through time:
            AA. Flag all records for that individual (set **consistencyTagIDTaxonIDQF** equal to 1), and
            BB. If there is a **taxonID** that is associated with the majority of records, assign all records involving that **individualID** to that **taxonID**, and re-assign values for **scientificName** and **taxonRank** (as in step 5.1.5.5).
            CC. Else, proceed to next step.
      B. If **sex** is Null:
         I. Insert -1 into corresponding cell in **consistencyTagIDSexQF** field

II. Else, if **sex** changes through time:

AA. Flag all records for that individual (set **consistencyTagIDSexQF** equal to 1), and

BB. Check for values of **reproductiveCondition** that indicate a reproductive state for this individual (i.e., if values in testes and nipples and pregnant and vagina fields of mam_capturedata_pub are neither NULL nor 'nonscrotal', 'nonenlarged', 'neither').If evidence of reproduction can be found, assign all records for that individual to the sex indicated in the record indicating reproduction. If the conflicting records are both indicative of reproduction (e.g., one record indicates a scrotal male and the other record a pregnant female), make no changes and go to next step. Else, if there is a sex associated with the majority of records, assign all records for that individual to that value of sex. Else, make no changes.

ii. Check for consistency and logical ordering of **lifeStage** designations for each individualID from one capture instance to the next

A. If **lifeStage** is Null:

I. Insert -1 into corresponding cell in **consistencyTagIDLifeStageQF** field

II. Insert -1 into corresponding cell in **orderTagIDLifeStageQF** field

B. Else, if **lifeStage** changes through time:

I. Set **consistencyTagIDLifeStageQF** to '1'

II. Check that the change follows a logical sequence, with the juvenile stage preceding the subadult stage (S) until the individual matures to an adult. Note that the intermediate Subadult stage is not necessarily captured, depending on the time elapsed between recapture events. Set **orderTagIDLifeStageQF** equal to 0 or 1, according to results of this test.

8. Remove records for species of conservation concern

a. For each record in *mam_capturedata_pub*:

i. Use taxonID and domainID to lookup the dXXNativeStatus for the relevant domain (e.g., d10NativeStatus where domainID=D10), based on AD[15]

ii. If dXXNativeStatus = 'STATE ' or 'FEDERAL':

A. Remove record from mam_capturedata_pub.

6 UNCERTAINTY

There are multiple sources of uncertainty associated with the small mammal capture data: error associated with plot establishment; error associated with setting of traps, in terms of geographic location and/or number of traps; human error in assignment of trapCoordinates; field measurement error associated with measuring and handling a live, mobile subject; potential for misidentification of species by parataxonomists and of age, sex and reproductive status, given the oft subjective nature of these determinations. Moreover, the vagaries of trapping frequently produce traps that are closed, disturbed, or destroyed (and therefore unavailable) but contain no captures, traps that are open but contain sign of a visiting small mammal, and occasional mortalities due to trapping. All of these factors combine to increase the uncertainty in aggregate data products, such as density estimates and capture rates.

In addition to the sources of uncertainty described above, traps used to capture animals may bring about biases that relate to their design, given that there is variation in: attractiveness of the bait and/or trap design to individuals of different sexes or life stages within and across species (Wilson et al. 1996), and the size of animals most likely to set off the triggering mechanism for a particular trap. These factors potentially result in biased estimates of relative abundances across sex, age, and species.

6.1 Reported Uncertainty

Although no quantitative algorithms are available to incorporate many of these sources of uncertainty into the associated data products, NEON can produce summary uncertainty reports for observational data products. The algorithm for assessing error rates in taxonomic identifications is presented in the NEON Algorithm Theoretical Basis Document: Taxonomic Consensus Identifications and Uncertainty [AD[07]]. The summary reports will include: bout-level data entry error rates; automated QA/QC error rates for each data product per bout; proportion of traps that have trapStatus values impacting trap availability; and the precision of the measurement tools. The rulers used for linear measurements provide measurements to the nearest millimeter. The spring scales used for weighing small mammals vary in precision: for individuals <100 grams, weight will be reported to the nearest gram; for individuals >100 grams, the spring scale has an uncertainty of +- 10 grams.

7 VALIDATION AND VERIFICATION

7.1 Algorithm Validation

7.2 Data Product Validation

7.3 Data Product Verification

8 SCIENTIFIC AND EDUCATIONAL APPLICATIONS

NEON small mammal sampling will assess the abundance and diversity of the nocturnal small mammal communities of North America, including the population dynamics of species that are competent reservoirs for infectious disease, at up to 60 sites for a period of 30 years. This represents a significant increase in the number and diversity of long-term datasets involving small mammals, and will also be unique in the degree of standardization across studies, as well as the availability of the data and archived samples to the scientific community and the public. This will allow NEON and the scientific community to address a diversity of questions (Table 2), and the associated vouchering of specimens and tissue samples will provide critical resources for external Principal Investigator-driven research to address an even wider range of questions.

Table 2. Examples of science questions that could be addressed with NEON data.

How do small mammal communities vary both within core sites and across land use types and ecoregions?
Which bioclimatic and habitat factors best predict the species composition of small mammal communities?
How do climate-driven plant and insect resources determine the population growth, fecundity, and density of small mammal populations?
How do changes in biodiversity affect resource use and infectious disease dynamics?
What are the specific local habitat traits (e.g., vegetation, slope, soil moisture, insect abundance etc.) favored by *Peromyscus spp.* that constitute refugia for hantavirus dynamics?
What is the time frame for the response of small mammal host populations to bioclimate driven resource pulses and what is the threshold density for increased hantavirus transmission?

9 FUTURE MODIFICATIONS AND PLANS

1. PDA interface:
    a. Referential integrity between plot night and trap night data
    b. Identical records test: warn the user if two records entered on a PDA/UI have the same values in all fields
2. Specific instructions for handling multiple sources of taxonomic identifications and generating a consensus identification along with categorical uncertainty designations will be contained in a separate ATBD that is currently in preparation (AD[07]).
3. Guidelines for uncertainty calculations and reporting for all data products are currently lacking, and need to be addressed, where possible.
4. Guidelines for releasing revised data products.
5. Future Tests:

a. Assign identificationStatus according to algorithm described in ATBD for QA/QC of Taxonomic classification of organismal observations [AD[07]].
    i. An algorithm that incorporates DNA barcode data from external laboratories, expert taxonomist determinations, and identificationQualifier values from the parataxonomist in the field to generate a consensus taxonID with an associated uncertainty value for a particular occurrence will ultimately be described in AD[07]. This algorithm will involve a more complicated decision tree and additional fields/tables of data products.

c. Incorporate Organismal Range Tests on the size measurements by species in either the PDA/webUI or the ATBD – data and associated algorithm still in development.

d. If the **tagID** is not found in any previous records:
    i. If **recapture** = 'Y':
        A. If length(**tagID**) >= 6 characters:
            I. Run a fuzzy matching algorithm (Modified Damerau-Levenshtein distance; Boehmer and Rees 2008) to check for possible errors in data transcription.
            II. If only one alternative ID is proposed:
                AA. If site, taxonID, and sex are the same in the current record as in previous records of the proposed tagID, the proposed tagID should be accepted as the correct value.
        B. Else: flag record (i.e., set **orderTagIDRecaptureQF** of the current record equal to 1).

10 BIBLIOGRAPHY

Boehmer, B., and T. Rees. 2008. The MDLD (Modified Damerau-Levenshtein Distance) Algorithm - TAXAMATCH. https://wiki.csiro.au/display/taxamatch/The+MDLD+(Modified+Damerau-Levenshtein+Distance)+Algorithm.

Efford, M. G., D. L. Borchers, and A. E. Byrom. 2009. Density estimation by spatially explicit capture-recapture: likelihood-based methods. Pages 255–269 *in* D. L. Thomson, E. G. Cooch, and M. J. Conroy, editors. Modeling demographic processes in marked populations. Springer US, Boston, MA.

Frase, B. A., T. M. Pizzuto, and L. L. Getz. 1990. Survivorship of bled voles measured by recapture. Journal of Mammalogy 71:104–105.

Gibson, C. M., R. H. Kao, K. K. Blevins, and P. D. Travers. 2012. Integrative taxonomy for continental-scale terrestrial insect observations. PLoS One 7:e37528.

Janzen, D. H., and W. Hallwachs. 2011. Joining inventory by parataxonomists with DNA barcoding of a large complex tropical conserved wildland in northwestern Costa Rica. PLoS One 6:e18123.

Kao, R. H., C. M. Gibson, R. E. Gallery, C. L. Meier, D. T. Barnett, K. M. Docherty, K. K. Blevins, P. D. Travers, E. Azuaje, Y. P. Springer, K. M. Thibault, V. J. McKenzie, M. Keller, L. F. Alves, E. S. Hinckley, J. Parnell, and D. S. Schimel. 2012. NEON terrestrial field observations: designing continental-scale, standardized sampling. Ecosphere 3:115.

Meeus, J. 1991. Astronomical Algorithms. Page 429. Willmann-Bell.

Merritt, J. F. 2010. The Biology of Small Mammals. Page 336 pp. The Johns Hopkins University Press, Baltimore.

Nagy, K. A., I. A. Girard, and T. K. Brown. 1999. Energetics of free-ranging mammals, reptiles, and birds. Annual Review of Nutrition 19:247–277.

Ostfeld, R. S., and R. R. Parmenter. 2008. Small mammals, Fundamental Sentinel Units TIGR team report.

Reed, A. W., and N. A. Slade. 2007. Demography and environmental stochasticity: empirical estimates of survival in three grassland rodents. Journal of Zoology 272:110–115.

Royle, J. A., J. D. Nichols, K. U. Karanth, and A. M. Gopalaswamy. 2009. A hierarchical model for estimating density in camera-trap studies. Journal of Applied Ecology 46:118–127.

Stapp, P. 2010. Long-term studies of small mammal communities in arid and semiarid environments. Journal of Mammalogy 91:773–775.

Thibault, K. M., S. K. M. Ernest, and J. H. Brown. 2010. Redundant or complementary? Impact of a colonizing species on community structure and function. Oikos 119:1719–1726.

White, G. C., and K. P. Burnham. 1999. Program MARK: Survival estimation from populations of marked animals. Bird Study 46 Supplem:120–138.

Wilson, D. E., F. R. Cole, J. D. Nichols, R. Rudran, and M. S. Foster (Eds.). 1996. Measuring and Monitoring Biological Diversity: Standard Methods for Mammals. Page 409 Biological Diversity Handbook Series. Smithsonian Institution Press, Washington, D.C.

11 APPENDIX: PDA & WEBUI DATA ENTRY AND VALIDATION REQUIREMENTS

11.1 Common User Interface Requirements

11.1.1 General Requirements

1. Follow guidelines as specified in the small mammal data ingest workbook: mamFieldSummary_PDA (AD[11]).
    a. When a new record is created follow guidelines for default values, as specified in **defaultValuePDA** and **defaultValueUI**.
    b. All fields must follow guidelines for **case**.
2. The user shall have the ability to delete any record which is not finalized. In order to delete a perPlotNight record, all perTrapNight records must first be deleted.

11.1.2 perPlotNight User Interface Requirements

1. Provide drop-down menu for domainID
2. Provide drop-down menu of siteID, filtered by the selected domainID
3. Provide drop-down menus of plotID, filtered by the selected siteID
    a. To generate menu of plotIDs, use the TOS plot-level spatial data lookup table (AD[14]) and where 'mam' is an element of **applicableModules**
4. Provide drop-down menu of available versions of samplingProtocol
5. identificationReferences should be a free-form text field (until a list can be compiled by Science)
6. Provide an 11 x 11 grid of buttons labeled with trap coordinates representing the 100 small mammal traps [A-J][1-10] as well as the unknown traps indicated by an 'X'. For example AX indicates a trap on column A with an unidentified row, 1X indicates a trap on row 1 with an unidentified column, and XX indicates a trap whose coordinate is completely unknown.
    a. Allow user to select each trap and set trapStatus options: 0 - no data; 1 - trap not set / captures(s) not processed; 2 - trap disturbed/door closed but empty; 3 - trap door open w/ spoor left; 4 - >1 capture in one trap; 5 – capture; 6 – trap set and empty
    b. The default setting for all traps will be no data
    c. Provide the ability to:
        i. Set the trapStatus for all remaining trapCoordinates to '6 – trap set and empty'
        ii. Set the trapStatus for all trapCoordinates to '1 - trap not set / capture(s) not processed'. This should only be allowed if here is no trap level-data.
    d. Resulting data sent to Maximo/PDR:
        i. A minimum of 100 trap coordinates with associated trap status data
        ii. No data for unknown coordinates (i.e., coordinates containing 'X') needs to be provided, unless intentionally entered by user

7. If either of the options '4->1 capture' or '5-capture' is selected, a new data entry form is provided to enter the perTrapNight data

11.1.3 perTrapNight User Interface Requirements

1. Auto-populate location (domainID, siteID, plotID, and trapCoordinate) and date based on values in the perPlotNight table
2. taxonID:
    a. lookup table: NEON.DOC.002259 NEON Taxonomic Name List for Small Mammals (AD[15])
    b. field is a typeahead in which the user can type in taxonID or scientificName
    c. populate list of values with taxa that are known to occur within the relevant domain (D##), using the D##NativeStatusCode field in the lookup table. Taxa to be included are indicated by all values in this field that are not 'A' for absent.

11.2 PDA User Interface Requirements

In order to minimize the amount of information stored on each PDA and bandwidth between the PDA and Maximo it was decided that only the 'perTrapNight' records entered would be sent to Maximo instead of sending a minimum of 100 'perTrapNight' records. To accommodate this change, the perPlotNight user interface is modified as follows:

1. Add an allTrapsEmpty field. Choices are 'no data', 'Yes', 'No'
    a. If 'Yes' any missing records will be generated with a trap status '6 – trap set and empty' will be stored in PDR with the rest of the captures.

2. Add an allTrapsNotSet field. This field will only be transmitted between the PDA and Maximo. Choices are 'no data', 'Yes', 'No'
    a. This field can only be set if there are no 'perTrapNight' records.
    b. If 'Yes' all 100 records (one for each trap) will be generated with a trap status '1 - trap not set' and stored in PDR with the rest of the captures.

11.2.1 perTrapNight User Interface Requirements

1. sampleIDs: User will scan the QR code to populate field
2. recordedBy – Defaults to the user logged in to the Mobile Data Recorder (MDR) app when the record is created. This field cannot be edited.

11.3 WEBUI User Interface Requirements

1. sampleIDs:
    a. Prompt the user to indicate whether a particular sample has been collected and therefore initiate the sampleID generation

i. UI: Concatenate: siteID + '.' + date + '.' + tagID + '.' + sampleTypeCode
   A. sampleTypeCodes:
      I. blood = B
      II. ear = E
      III. feces = F
      IV. hair = H
      V. whisker = H
         BB. *Note: hair and whiskers are stored in the same coin envelope and should therefore receive the same sampleID*
      VI. voucher = V
   B. Example: CPER.20120720.R1234.B

b. If the user indicates that a blood sample has been collected, the bloodSampleMethod should then be entered
c. If the user indicates that a fecal sample has been collected, the fecalSampleCondition should then be entered
d. If value in fate field is 'dead', prompt the user to choose whether or not to generate a voucher sampleID. Prompt the user to provide an explanation in the remarks if 'no' is selected.

2. measuredBy and recordedBy – Maximo user list for FOPs or 'Other'
   a. UI solution: TBD

11.4 Common Validation Requirements

All data must pass the field specific validation rules found in the small mammal data ingest workbook: mamFieldSummary_PDA (AD[11]).

The rules are to be evaluated in the following order:

1. Follow guidelines for fields in which no data have been entered, as specified in **noDataOutcomePDA (for mobile devices)** and **noDataOutcomeUI (for WEBUI)**:
   a. If **noDataOutcome** = fail:
      i. do not let user finalize record until a value is provided
      ii. warning message text, 'Please enter a value for [fieldname] to continue', unless an alternative is provided in **warningText**
   b. Else if **noDataOutcome** = warn:
      i. warn user that a value is missing prior to finalizing record, but allow selection of 'OK' to continue without a value
      ii. warning message text, 'Please confirm that there is no value for [fieldname] to continue', unless an alternative is provided in **warningText**

c. Else if **noDataOutcome** = pass:
    i. allow user to finalize record with no values in this field

2. If a field has data, constrain the entered value to the correct **dataType.** (If the field is null or empty only **noDataOutcome** is performed.) If the field has data and the data is not of the correct type display an error message stating 'The value for [fieldname] must be a [datatype]' and do not allow the record to be finalized.
3. If a field has data of the correct data type, constrain entered values to valid entries, as described in the **entryValidationRules** and do not allow the record to be finalized if it does not pass validation. Display an error message based on the type of error: For example, for range errors**,** display an error message stating 'The value for [fieldname] must be between [minval] and [maxval]'.

11.5 PDA Validation Requirements

11.5.1 perTrapNight Validation Rules

1. tagType / tagID:
    a. If tagType = 'Ear' then tagID must be in the following format: L####, R#### where # represents any number
    b. If tagType = 'Pit' then use Bluetooth-enable RFID reader to scan RFID and store entire serial number. No validation is performed other than there must be a value in tagID
    c. If tagType = 'Other' then tagID must start with an 'O' and be followed by 1 or more characters
    d. If tagType = 'None' then tagID must not contain a value.

11.6 WEBUI Validation Requirements

11.6.1 perTrapNight Validation Rules

1. tagID:
    a. Ear tags: L####, R####, or O*|XXXXX, where X is alphanumeric
        i. *Note: The latter option is for the uncommon event that an individual is captured that has been tagged by other researchers*
    b. PIT tag: last 6 digits of serial number