VectorPredictor Data Requirements and Standard Operating Proced

Document ID: VP-SOP-001
Version/date: 1.0 / 2026-04-27
Project: VectorPredictor, Royal Society International Collaboration Awards

1. Purpose

This SOP defines the minimum data requirements and standard workflow for preparing, uploading, quality-checking, and analysing mosquito mid-infrared or near-infrared spectral data in VectorPredictor. It is intended for laboratory users, field teams, data managers, and analysts using the web application to predict mosquito age and species.

2. Scope

This SOP applies to:

* Mosquito specimen metadata entry.
* Collection location records.
* Spectral data uploads in CSV, TXT, Excel, and DAT formats.
* Optional insecticide resistance bioassay records.
* Quality control checks and prediction result review.

It does not replace local laboratory safety procedures, ethics requirements, or instrument-specific maintenance instructions.

3. Responsibilities

* Sample collector: records accurate collection site, date, environmental conditions, species, sex, and physiological status.
* Laboratory operator: prepares specimens, operates the spectrometer, exports spectral files, and records measurement parameters.
* Data manager: checks filenames, metadata completeness, file format compliance, and data provenance before upload.
* Analyst: reviews quality indicators, runs predictions, interprets outputs, and records any exclusions or repeat measurements.
* Project lead or delegated reviewer: approves final datasets used for reporting, publication, or model evaluation.

4. Required Accounts and System Access

1. Use a modern browser such as Chrome, Firefox, Safari, or Edge.
2. Create a VectorPredictor user account with a valid username, email address, and secure password.
3. Log in before creating locations, metadata, spectra, bioassays, or prediction records.
4. Use one account per named user where possible so that records remain attributable.

5. Data Requirements

5.1 Specimen Metadata

Each spectral upload must be linked to a metadata record. The following fields are required in VectorPredictor:

* Metadata name: a short unique identifier for the sample or batch.
* Spectrometer: brand and model of the instrument.

* Min wavenumber: lower measured wavenumber, accepted range 100-10000 cm-1.
* Max wavenumber: upper measured wavenumber, accepted range 100-10000 cm-1.
* Scans per spectrum: number of scans averaged, accepted range 1-128.
* Spectral resolution: value and units, for example 1 cm-1 or 4 cm-1.
* Body part: body part scanned, for example head/thorax or abdomen.
* Origin: Laboratory, Field-collected, or Unknown.
* Collection location: existing map location record.
* Collection date: date collected in YYYY-MM-DD format.
* Storage temperature: accepted range -80 to 50 deg C.
* Collection temperature: accepted range -20 to 60 deg C.
* Relative humidity: accepted range 0-100%.
* Physiological status: Blood fed, Partially blood fed or semigravid, Sugar fed or unfed, Gravid, or Unknown.
* Sex: Female or Male.
* Species: selected from the species list.

5.2 Location Data

Create location records before metadata entry. Required location information is:

* Location name.
* Coordinate selected on the map or decimal latitude and longitude where supported.
* Optional description, such as village, insectary, trap site, or collection notes.

Latitude must be within -90 to 90 decimal degrees. Longitude must be within -180 to 180 decimal degrees.

5.3 Spectral Data Files

Supported file types are:

* CSV: comma-separated values.
* TXT: tab-separated, space-separated, or semicolon-separated values.
* Excel: XLSX or XLS, with data on the first sheet.
* DAT: tab-separated spectral records with metadata columns and 1625 spectral values.

General spectral requirements:

* Files must contain numeric spectral measurements.
* Files must have at least two columns for standard wavenumber and absorbance/intensity data, or 1625 feature columns for model-ready matrices.
* Wavenumbers should usually cover 400-4000 cm-1; the application accepts 100-10000 cm-1 but unusual ranges must be reviewed.
* Recommended resolution is 1-4 cm-1.
* Recommended data volume is at least 1000 spectral points for analytical use. Demonstration files may contain fewer points but should not be treated as production data.
* Absorbance/intensity values should normally be finite numeric values, typically 0.0-2.0 for absorbance unless baseline correction creates another valid range.
* Missing values, text within numeric columns, non-finite values, corrupted files, and unlabelled data should be corrected before upload.
* Text files should use UTF-8 where possible.
* Recommended file size is below 5 MB; maximum file size is 10 MB per upload.

DAT file requirements:

* Each spectrum must be on one line.
* Each valid line must start with AA.
* The first 9 columns contain metadata values.
* The next 1625 columns contain numeric spectral intensity values.

* The file must be tab-separated.

5.4 Bioassay Data, If Available

Bioassay records are optional but should be linked to the relevant metadata record when insecticide resistance information is available. Required fields are:

* Insecticide name.
* Assay type.
* Susceptible: Yes or No.
* Resistant: Yes or No.
* Unknown: Yes or No.
* Associated metadata record.

Only one resistance interpretation should be selected as the primary state wherever possible. If results are ambiguous, record Unknown and add explanatory notes outside the application if needed.

6. Pre-Upload Quality Checks

Before uploading files:

1. Confirm that the metadata record has been created and reviewed.
2. Confirm that species, sex, body part, origin, collection date, and location are correct.
3. Verify the spectral file extension matches the actual file format.
4. Open the file locally and confirm that numeric data are present.
5. Check that wavenumbers are in cm-1 and are consistently ordered.
6. Remove unrelated header/footer text, comments, blank rows, and merged cells.
7. Confirm there are no missing values in spectral columns.
8. Use clear filenames containing sample ID, date, instrument, and replicate where appropriate.
9. Retain a read-only copy of the raw instrument export before any preprocessing.

Recommended filename pattern:

```
site_sampleid_species_bodypart_YYYYMMDD_replicate.extension
```

Example:

```
hut05_agambiae_headthorax_20230819_rep01.csv
```

7. Standard Operating Procedure

7.1 Create or Verify Reference Records

1. Log in to VectorPredictor.
2. Add any missing mosquito species records.
3. Add any missing body part records.
4. Create or verify the collection location record.
5. Confirm that the location coordinates and description are correct.

7.2 Create Metadata

1. Navigate to the metadata creation page.
2. Enter all required specimen, environmental, and spectral measurement fields.
3. Use the same units and naming conventions across a dataset.
4. Submit the metadata form.
5. Review the saved metadata record before proceeding.

7.3 Prepare Spectral Files

1. Export spectra from the instrument in CSV, TXT, XLSX, XLS, or DAT format.
2. Preserve the raw export in a secure project folder.
3. If preprocessing is required, document the method used, such as baseline correction, smoothing, normalisation, or peak alignment.
4. Confirm that the final upload file meets the requirements in section 5.3.
5. If using CSV, TXT, or Excel with two columns, ensure the columns represent wavenumber and absorbance/intensity.
6. If using DAT, confirm that each spectrum has 1625 spectral values after the first 9 metadata columns.

7.4 Upload Spectra

1. Navigate to the spectra upload page.
2. Select the correct metadata record.
3. Choose one or more spectral files.
4. Upload the file or files.
5. Wait until processing completes.
6. Record any upload errors and correct the source file rather than forcing an upload.

7.5 Review Uploaded Spectra

1. Open each uploaded spectrum or batch result.
2. Review the spectral plot for obvious artefacts, flat lines, saturation, excessive noise, or unexpected gaps.
3. Review quality filter outputs.
4. Confirm the raw values displayed by the system match the uploaded file.
5. Flag or exclude any spectrum that fails quality review.

7.6 Run Predictions

1. Navigate to the prediction interface.
2. Select the metadata record or uploaded spectrum.
3. Select the appropriate model if multiple models are available.
4. Confirm that body part, species context, and file format are compatible with the intended model.
5. Run the prediction.
6. Review predicted age, species prediction, confidence scores, and any warnings.
7. Export or record results according to the project data management plan.

7.7 Batch Processing

VectorPredictor can process multiple spectra in a single supported file. For batch uploads:

1. Confirm that every row or line represents one valid spectrum.
2. Ensure all spectra in the file share the same metadata context, or split them into separate files before upload.
3. Review individual results for each spectrum; do not approve an entire batch solely because the upload succeeded.
4. Investigate any spectrum-specific errors before finalising the dataset.

8. Automated Quality Indicators

The application may apply checks including:

* Low intensity filter: checks whether signal intensity in key regions is above threshold.
* Abnormal background filter: checks for unusual background behaviour.
* Atmospheric interference filter: checks for potential water or carbon dioxide interference.

* Data completeness checks: detect empty files, missing values, and non-numeric content.
* Format compliance checks: confirm that the file structure can be parsed.
* Wavenumber range warnings: identify values outside the expected analytical range.

Failed quality checks should be reviewed by the analyst. Common corrective actions include re-exporting the file, removing formatting artefacts, re-measuring the specimen, checking background correction, and confirming instrument settings.

9. Acceptance Criteria

A dataset is acceptable for prediction when:

* Required metadata fields are complete and internally consistent.
* The uploaded file uses a supported format.
* Spectral values are numeric and finite.
* The wavenumber range is plausible for the instrument and model.
* The spectral plot shows a biologically plausible, non-corrupted spectrum.
* Quality filters pass or any failures are reviewed and justified.
* The metadata record matches the specimen or batch represented by the uploaded spectra.

10. Rejection Criteria

Reject or rework a spectrum before analysis when:

* The file cannot be parsed.
* Required metadata are missing or linked to the wrong specimen.
* Numeric columns contain text, missing values, or non-finite values.
* The spectrum has obvious instrument failure, saturation, truncation, or extreme noise.
* The wavenumber range is incompatible with the selected model.
* DAT files do not contain 1625 spectral values per spectrum.
* The batch contains mixed specimens that require separate metadata records.

11. Data Security and Governance

* Upload only data approved for use in the project.
* Avoid including personal information in filenames, free-text descriptions, or metadata fields.
* Keep original instrument files in a controlled project location.
* Maintain versioned records of any preprocessing steps.
* Do not share user credentials.
* Review exported prediction results before external sharing.
* Follow local institutional policies for data storage, transfer, and publication.

12. Troubleshooting

Invalid file format:

* Confirm that the extension is CSV, TXT, XLSX, XLS, or DAT.
* Re-export from the instrument or convert to CSV.
* Check that the file is not corrupted.

No numeric data found:

* Remove text columns from the spectral data area.
* Confirm that decimal separators are compatible with the parser.
* Check that the first sheet of an Excel workbook contains the data.

Wavenumber range warning:

* Confirm units are cm-1.
* Check that the wavenumber column has not been replaced by row numbers.
* Confirm min and max wavenumber metadata.

Low intensity or poor background:

* Check background correction.
* Confirm the specimen was positioned correctly.
* Inspect water and carbon dioxide interference.
* Re-measure the specimen if required.

Prediction failed:

* Confirm that the file has either two spectral columns or 1625 model-ready features.
* Confirm that all values are finite.
* Try a single known-good sample file to distinguish data issues from system issues.

13. Minimum Upload Checklist

Use this checklist before uploading production data:

* User is logged in with the correct account.
* Species and body part reference records exist.
* Collection location is accurate.
* Metadata record is complete.
* Spectrometer, wavenumber range, scan count, and resolution are recorded.
* Environmental conditions are within accepted ranges.
* Spectral file uses a supported format.
* File contains numeric values only in spectral columns.
* File size is below 10 MB.
* Raw file has been backed up.
* Any preprocessing has been documented.
* Uploaded spectra have been visually reviewed.
* Failed quality checks have been investigated.