# GENHackX: Program Overview and In-Depth Methods

November 17, 2025

**Abstract**

The GENHackX repository provides a comprehensive suite of Python scripts for advanced geospatial data analysis. This document outlines the primary tools within the repository, which focus on spatial resolution assessment, metadata extraction, and data alignment for heterogeneous datasets including Sentinel-2, Sentinel-3, ERA5, and GPKG vector files. We provide a detailed mathematical development of the core statistical methods used in our resolution analysis: "Variance Explained" and "Moran's I" for spatial autocorrelation. These methods form a robust framework for identifying the optimal spatial resolution that balances data fidelity with computational efficiency.

## 1 Overview

The GENHackX repository contains a suite of Python scripts for geospatial data analysis. The primary focus is on assessing spatial resolution, extracting critical metadata, and performing data alignment for diverse datasets such as Sentinel-2, Sentinel-3, ERA5, and GPKG vector files. This document details the available scripts and the statistical methodologies underpinning their analysis functions.

## 2 Script Descriptions

- **data_resolutions.py**: Extracts spatial metadata (CRS, bounds, feature count, and estimated resolution) from GPKG, Sentinel-2, Sentinel-3, and ERA5 NetCDF files. Outputs a summary of spatial characteristics for all datasets.
- **test_resolution_framework.py**: Provides a framework to evaluate the effect of changing spatial resolution on raster datasets. For each candidate resolution, it computes:
  - **Variance Explained**: Quantifies how much of the original data's variance is preserved after aggregation to a coarser grid.
  - **Moran's I**: Measures spatial autocorrelation, indicating the degree to which similar values cluster spatially at each resolution.

  This script helps identify the coarsest resolution that retains the essential spatial characteristics of the data.
- **alignment.py**: Contains functions for aligning and reprojecting raster datasets to a common grid, ensuring spatial comparability across sources.
- **inspect_gpkg.py**: Inspects GPKG vector files, listing layers, columns, and previewing attribute data for exploratory analysis.
- **read_docx.py**: Reads and extracts text from Microsoft Word (.docx) files for documentation or metadata review.

# 3 Mathematical Methods Explained

This section provides a more detailed development of the statistical methods used in the `test_resolution_frame` script.

## 3.1 Variance Explained

The "Variance Explained" metric is used to quantify the amount of information loss that occurs when a high-resolution raster is aggregated to a coarser resolution. It is analogous to the coefficient of determination ($R^2$) in a regression context.

Let $X$ be the original, high-resolution raster, represented as a set of $N$ cell values $\{x_1, x_2, \ldots, x_N\}$. Let $\hat{X}$ be the corresponding raster after being aggregated (downsampled) to a coarser resolution and then upsampled back to the original resolution for comparison. The values of $\hat{X}$ are $\{\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_N\}$.

**1. Total Variance** First, we define the total variance of the original data, which represents the total information or variability present. This is based on the Total Sum of Squares (TSS):

$$TSS = \sum_{i=1}^{N} (x_i - \bar{x})^2 \tag{1}$$

where $\bar{x}$ is the mean of the original data. The variance is then:

$$\text{Var}(X) = \frac{TSS}{N-1} \tag{2}$$

**2. Residual (Error) Variance** Next, we define the residual, or error, $E$, as the cell-wise difference between the original data and its resampled version:

$$E = X - \hat{X} \tag{3}$$

The variance of this residual, $\text{Var}(E)$, quantifies the variability *not* captured by the coarser-resolution model. This is based on the Residual Sum of Squares (RSS):

$$RSS = \sum_{i=1}^{N} (e_i - \bar{e})^2 = \sum_{i=1}^{N} \left( (x_i - \hat{x}_i) - \overline{(x - \hat{x})} \right)^2 \tag{4}$$

Assuming the aggregation method is unbiased (i.e., the mean of the errors $\bar{e}$ is zero), this simplifies to $RSS \approx \sum_{i=1}^{N} (x_i - \hat{x}_i)^2$. The residual variance is:

$$\text{Var}(E) = \text{Var}(X - \hat{X}) = \frac{RSS}{N-1} \tag{5}$$

**3. Final Formulation** The "Variance Explained" (VE) is the proportion of the original variance that is *not* residual variance. It is calculated as:

$$\text{VE} = 1 - \frac{\text{Var}(E)}{\text{Var}(X)} = 1 - \frac{RSS/(N-1)}{TSS/(N-1)} \tag{6}$$

This simplifies to the final form, which is identical to $R^2$:

$$\text{Variance Explained} = 1 - \frac{\sum_{i=1}^{N} (x_i - \hat{x}_i)^2}{\sum_{i=1}^{N} (x_i - \bar{x})^2} \tag{7}$$

A value close to 1 indicates that the aggregated raster $\hat{X}$ preserves most of the original data's variability. A sharp drop in VE as resolution coarsens indicates a significant loss of spatial detail.

### 3.2 Moran's I (Spatial Autocorrelation)

Moran's I is a measure of global spatial autocorrelation. It assesses whether the spatial distribution of a variable is clustered, dispersed, or random.

The formula requires several components: the variable's deviation from its mean, a spatial weights matrix, and scaling factors.

**1. Spatial Weights Matrix ($w_{ij}$)** The core of any spatial autocorrelation statistic is the spatial weights matrix, $W$. This $N \times N$ matrix (where $N$ is the number of spatial units) defines the neighborhood relationship between all pairs of locations $i$ and $j$.

- $w_{ij} > 0$ if $i$ and $j$ are considered neighbors.
- $w_{ij} = 0$ if they are not neighbors.
- By convention, $w_{ii} = 0$.

Common definitions for $w_{ij}$ include contiguity (1 if $i$ and $j$ share a border, 0 otherwise) or inverse distance. The total sum of all weights is $W = \sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij}$.

**2. Components of the Formula** Moran's I is essentially a spatially weighted correlation coefficient. We can deconstruct its formula:

$$I = \frac{N}{W} \frac{\sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

Let's analyze the parts:

- $\sum_{i=1}^{N}(x_i - \bar{x})^2$: This is the Total Sum of Squares (TSS), measuring the overall variance of the data.
- $\sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij}(x_i - \bar{x})(x_j - \bar{x})$: This is the spatially weighted covariance term. For each pair of locations $(i, j)$, it multiplies their respective deviations from the mean. This product is positive if both locations are above the mean or both are below the mean. The term is then scaled by the spatial weight $w_{ij}$. If $i$ and $j$ are not neighbors ($w_{ij} = 0$), they do not contribute to the sum.
- $\frac{N}{W}$: This is the normalization constant. $N$ is the number of spatial units, and $W$ is the sum of all weights.

**3. Interpretation** The value of Moran's I generally ranges from -1 to +1.

- $I > 0$: Indicates positive spatial autocorrelation (spatial clustering). Similar values (high-high or low-low) are clustered together.
- $I < 0$: Indicates negative spatial autocorrelation (spatial dispersion). Similar values are far apart, resembling a checkerboard pattern.
- $I \approx 0$: Indicates spatial randomness (no autocorrelation). The expected value under the null hypothesis of no autocorrelation is $E(I) = -1/(N-1)$, which approaches 0 for large $N$.

By computing Moran's I at different resolutions, the script assesses how this fundamental spatial structure is preserved or lost as the grid is coarsened.