

**Institute of Primate Research**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **SOP No.** | **Issue Number** | **Issue Date** | **Revision Status** | **Revision Date** |
| **SOP/KIPRE/RPD/DSAS/3.1.76** | **Version 01** | **October 2025** | **-** | **-** |

**STANDARD OPERATING PROCEDURE (SOP) DOCUMENT**

**Handling large datasets and trend detection**

|  |  |  |  |
| --- | --- | --- | --- |
| **Approvals** |  |  |  |
|  | **Name** | **Signature** | **Date** |
| **Developed by:** | \_Patrick Waweru Mwaura\_ | **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_** | **\_6th October; 2025\_** |
|  | \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ | **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_** | **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_** |
|  | \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ | **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_** | **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_** |
| **Reviewed by:** | **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_** | **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_** | **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_** |
| **Approved by:** | **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_** | **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_** | **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_** |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |

**Table of Contents**

[1. PURPOSE 4](#_Toc144316958)

[2. SCOPE 4](#_Toc144316959)

[3. PERSONS RESPONSIBLE: 4](#_Toc144316960)

[4. FREQUENCY 4](#_Toc144316961)

[5. MATERIALS 4](#_Toc144316962)

[6. PROCEDURE 4](#_Toc144316963)

[7. REFERENCES 7](#_Toc144316964)

# 

# PURPOSE

To standardize DS&AS methods for managing large datasets and detecting temporal, spatial, and epidemiological trends.

# SCOPE

Applies to DS&AS projects involving large-scale biomedical, ecological, genomic, or public health data.

# PERSONS RESPONSIBLE:

* **Data Engineer:** Manages big data storage and access pipelines.
* **Data Scientist:** Conducts statistical and trend analysis.
* **Head of DS&AS:** Ensures compliance and efficiency.

# FREQUENCY

* Applied continuously for projects requiring high-volume or real-time analytics.
* Reviewed **annually** for optimization and scalability.

# MATERIALS

* Big data tools (Hadoop, Spark, SQL, NoSQL, and PostgreSQL).
* Trend detection tools (time-series models, anomaly detection, and geospatial analysis).
* Data visualization dashboards (R Shiny, Tableau, PowerBI).

# PROCEDURE

1. **Data Ingestion:** Import datasets into scalable storage (SQL/NoSQL databases, distributed systems).
2. **Preprocessing:** Apply automated cleaning, deduplication, and standardization.
3. **Trend Analysis:** Apply statistical methods (ARIMA, Cox models, GAMs) and ML-based detection (anomaly detection, clustering).
4. **Visualizations:** Generate dashboards for real-time monitoring of patterns and anomalies.
5. **Archiving:** Store processed datasets and scripts in the central repository.
6. **Review:** Conduct annual scalability and performance audits.

# REFERENCES