Institute of Primate Research

STANDARD OPERATING PROCEDURE (SOP) DOCUMENT

Handling large datasets and trend detection

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

Approvals

	Name	**Signature**	**Date**
Developed by:	_Patrick Waweru Mwaura_	_____	**_6th October; 2025_**
	_____	_____	_____
	_____	_____	_____
Reviewed by:	_____	_____	_____
Approved by:	_____	_____	_____

Table of Contents

1. PURPOSE

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

2. SCOPE

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

3. 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.

4. FREQUENCY

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

5. 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).

6. 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.

7. REFERENCES