# A computational framework to systematize uncertainty analysis in the sediment fingerprinting approach using least square methods

Lidiane Buligon, Tiago Martinuzzi Buriol, Jean Paolo Gomes Minella, Olivier Evrard

Computational and Applied Mathematics

# A computational framework to systematize uncertainty analysis in the sediment fingerprinting approach using Least Square Methods

--Manuscript Draft--

| | |
|---|---|
| **Manuscript Number:** | |
| **Full Title:** | A computational framework to systematize uncertainty analysis in the sediment fingerprinting approach using Least Square Methods |
| **Article Type:** | Original Article |
| **Funding Information:** | CEA-Atomic Energy Commission — Mr Olivier Evrard<br>National Council for Scientific and Technological Development (200008/2023-4) — Mrs Lidiane Buligon<br>CAPES-Coordination for the Improvement of Higher Education Personnel (88887.310201/2018-00) — Mrs Lidiane Buligon |
| **Abstract:** | Simulating sediment transfer processes in catchments has contributed significantly to solving environmental problems due to its importance in the silting of rivers and reservoirs and for controlling the pollution of water bodies. Among the methods used to improve data collection and modelling, the “sediment fingerprinting approach” uses tracers reflecting the composition of eroded soils and sediments in multivariate statistical analyses and mathematical models for optimizing equation systems. Based on generalized least squares (GLS) method and Mahalanobis distance, this study sought to present a computational framework to solve over-determined systems applied to sediment tracing, systematize the uncertainty analysis and sample number optimization. Hence, this approach takes into account the influence of collinearity among the chemical variables that compose the tracer set to be evaluated by the presence of the variance-covariance matrix. A dataset from the Arvorezinha experimental catchment in southern Brazil was used to validate the modeling, and our findings confirmed the assumption of increased uncertainty as the number of target samples decreases in the sources or eroded sediment samples. Sharing the file with the Python code contributes to improving the technique as it allows other researchers to systematically improve the definition of the number of samples required based on the uncertainty analysis. |
| **Corresponding Author:** | Lidiane Buligon, Dr<br>Federal University of Santa Maria: Universidade Federal de Santa Maria<br>Santa Maria, Rio Grande do Sul BRAZIL |
| **Corresponding Author Secondary Information:** | |
| **Corresponding Author's Institution:** | Federal University of Santa Maria: Universidade Federal de Santa Maria |
| **Corresponding Author's Secondary Institution:** | |
| **First Author:** | Lidiane Buligon, Dr |
| **First Author Secondary Information:** | |
| **Order of Authors:** | Lidiane Buligon, Dr<br>Tiago Martinuzzi Buriol, Dr<br>Jean Paolo Gomes Minella, Dr<br>Olivier Evrard, PhD |
| **Order of Authors Secondary Information:** | |
| **Author Comments:** | |
| **Suggested Reviewers:** | Gabriel Haeser<br>ghaeser@ime.usp.br |

| | |
|---|---|
| | Research Interests: Nonlinear Optimization, Conic Optimization, Numerical Analysis, Optimality Conditions, Algorithms, Applications. |
| | Patrick Laceby<br>patrick.laceby@gov.ab.ca<br>Skills and expertise:<br>Sediments, Geochemical Modeling Stable Isotopes, Sedimentology, Geochemistry Stable, Isotope Analysis |
| | Arman Haddadchi<br>aman.haddadchi@niwa.co.nz<br>Skills and expertise:<br>Genetic Algorithm, Soil Erosion ,Sediment, Sediments River, Engineering Environment ,Geomorphology |
| | Leonardo Ramos Emmendorfer<br>leonardo.emmendorfer@gmail.com<br>Skills and expertise: Evolutionary Computation, Clustering, Evolutionary Algorithms, Heuristics, Optimization |

# A computational framework to systematize uncertainty analysis in the sediment fingerprinting approach using Least Square Methods

Lidiane Buligon[1*], Tiago Martinuzzi Buriol[1], Jean Paolo Gomes Minella[2], Olivier Evrard[3†]

[1*]Department of Mathematics, Federal University of Santa Maria, Roraima Av., n.1000, Santa Maria, 97105-900, Rio Grande do Sul, Brazil.
[2]Department of Soils, Federal University of Santa Maria, Roraima Av., n.1000, Santa Maria, 97105-900, Rio Grande do Sul, Brazil.
[3]Laboratoire des Sciences du Climat et de l'Environnement (LSCE-IPSL), CEA Saclay, Orme des Merisiers , Gif-sur-Yvette, 91 191 Cedex, France.

*Corresponding author(s). E-mail(s): buligon.l@ufsm.br;
Contributing authors: tiago.buriol@ufsm.br; jean.minella@ufsm.br; olivier.evrard@lsce.ipsl.fr;
† These authors contributed jointly to this work

**Acknowledgments**

The authors acknowledge the financial support granted by the CNPq (National Council for Scientific and Technological Development) and CAPES (Coordination for the Improvement of Higher Education Personnel) by the first author's scholarship received during the development of this study. CEA (Atomic Energy Commission) in supporting a 1-year research visit in France. We would also like to thank Atlas Assessoria Linguística for support with the English version of this manuscript.

**Abstract**

Simulating sediment transfer processes in catchments has contributed significantly to solving environmental problems due to its importance in the silting of rivers and reservoirs and for controlling the pollution of water bodies. Among

the methods used to improve data collection and modelling, the "sediment fingerprinting approach" uses tracers reflecting the composition of eroded soils and sediments in multivariate statistical analyses and mathematical models for optimizing equation systems. Based on generalized least squares (GLS) method and Mahalanobis distance, this study sought to present a computational framework to solve over-determined systems applied to sediment tracing, systematize the uncertainty analysis and sample number optimization. Hence, this approach takes into account the influence of collinearity among the chemical variables that compose the tracer set to be evaluated by the presence of the variance-covariance matrix. A dataset from the Arvorezinha experimental catchment in southern Brazil was used to validate the modeling, and our findings confirmed the assumption of increased uncertainty as the number of target samples decreases in the sources or eroded sediment samples. Sharing the file with the Python code contributes to improving the technique as it allows other researchers to systematically improve the definition of the number of samples required based on the uncertainty analysis.

**Keywords:** Computational Mathematics, Generalized Least Squares, Sediment Source Identification, Mahalanobis Distance, Confidence Region

# 1 Introduction

Water erosion is one of the main processes of soil degradation, severely impacting water resources [1]. Researchers have demonstrated that modelling sediment production and transfer requires a set of mathematical techniques that rely on robust monitoring network [2]. The modelling techniques depend on the efficiency of the measured data in representing the spatial variability of sediment sources. In addition, the models seek to represent the temporal variability of eroded sediment mobilization and transfer phenomena. Thus, the ability of the models to represent the spatial and temporal variability of the phenomena depends essentially on the data sets available. On the other hand, the model outputs should maximize the explanatory capacity of the sample set, minimizing the costs associated with data collection and analysis. The sediments in the river originate from various locations where the erosion process can occur with specific patterns and magnitudes [3]. Therefore, quantitatively defining the contribution of each source is crucial to propose effective soil conservation measures and reduce environmental and economic problems linked to soil erosion.

Sediment source identification is an important branch of sediment production modeling that employs tracers found in soils and sediments in a set of statistical analysis techniques known as the "sediment source identification" or "sediment fingerprinting/tracing approach" [4–10]. This approach involves various areas of knowledge, including geoscience, statistics, and computational mathematics, fostering the exchange of knowledge from the choice of potential tracers to the use of mathematical and statistical methods to describe the different steps of the process. Optimizing models based on observing the processes improves the estimation of contributions from different sediment sources, quantifying the associated uncertainties and extracting the

maximum amount of information from the data set obtained in the field. Furthermore, the shared use of open-source algorithms expands the applications of these new methods among different research groups, thereby enhancing the scientific and technological advances in soil science and related disciplines. Numerous contributions have been made to improving techniques for identifying sediment sources for monitoring and modelling [3, 11–17]. The main assumptions of the suspended sediment identification technique are that: (1) sources can be discriminated by different characteristics (tracer properties) found in the source soils; (2) the eroded sediments in the river system consist of a mixture of the sediments originating from potential sources accumulated during the transfer process across the catchment; (3) the temporal variations in the tracer properties of the eroded sediments found in the river directly reflect the spatial variation of the erosion processes according to the behavior of each source; (4) the tracer properties of the sources and suspended sediments can be compared to establish the contribution of each source to the sampled sediment (target). Given these assumptions, the sediment fingerprinting approach establishes a relationship between source characteristics and those of suspended sediments by mathematically solving an over-determined linear system of equations (i.e., linear systems in which the number of equations is greater than the number of unknowns).

Applications in different areas of research are described by over-determined linear systems, including the traditional modeling of sediment source identification proposed by Walling and Woodward [18]. Methods for solving this type of system of equations have been continuously improved in order to obtain better estimates for the set of observed quantities. It is expected that the more information available, the better the quality of the results obtained from your analysis. However, obtaining feasible solutions for over-determined systems is a challenge. Regression analysis has been the most used technique in these cases, as it measures the direction and intensity of the relationship between the dependent and independent variables and numerically describes this relationship [19–22]. The methods use the mean square error to calculate and evaluate the performance of an estimator. In the ordinary least squares (OLS) method, a system of linear equations corresponds to a matrix equation of the form $Ax = b$, where the matrix $A$ and the vector $b$ are given, and $x$ is the unknown solution [23–25]. The OLS algorithm was employed to create the FingerPro package [26], which was developed in the R language to determine the contribution of sediment sources to target material. However, when there is a certain degree of correlation between the residuals in a regression model, the Aitken estimator (or Gauss-Markov estimator) should be applied. In practice, however, the covariance matrix of the error is generally unknown, making this estimator nonviable. In such cases, a Generalized Least Squares (GLS) estimator, which is defined as the Gauss-Markov estimator with the unknown covariance matrix replaced with a suitable estimator, is used [21]. Least square methods can lead to the obtainment of approximate solutions of over-determined systems, although sometimes no exact solution can be found. The basis for obtaining solutions to these types of systems is based on matrix algebra, which computational advances have been greatly boosted during the last several decades [27–29].

According to Walling [4] one of the challenges in improving the sediment source identification technique is related to the estimation of the uncertainties of the results.

From this perspective, the influence of the number of samples used to characterize both the potential sources and the target sediment is critical for evaluating the confidence level associated with the results. The assumption, in this case, is based on estimating the increase in uncertainty when the number of samples decreases. In addition, the number of tracers selected and their relationship with the number of sources are also topics under investigation [2]. Latorre et al. [13] emphasized that employing more tracers than the number of potential sources can lead to mathematical inconsistencies due to obtaining multiple solutions in over-determined systems. Although solutions of over-determined systems provide a possible source of errors for the sediment fingerprinting approach, the authors believe that this topic has remained under-investigated based on the premise that the selection of tracers and the potential advantage provided by the use of different models (e.g., FingerPro [26], MixSIAR [30], and SIFT [31]) have led to the primary advances of the technique.

In other studies, Haddadchi et al. [32] and Haddadchi et al. [33] compared the accuracy of different mixing models relying on the same source and target sediment data set. The results indicated that the relative contribution of sources to sediments depend on model was applied. In turn, Latorre et al. [13] and Lizaga et al. [12] pointed out that the use of different tracer selection methods may affect equally the outputs of the models. However, when using the same selection process, both types of models produced similar output results. Both studies agreed that the models based on FingerPro and MixSIAR did not test for consistency and the existence of multiple solutions in over-determined systems; for this reason, they recommend the implementation of new tracer selection method. In this context, Latorre et al. [13] designed the Consistent Tracer Selection (CTS) method to extract the solutions in the datasets of each mixture (from the selection of subsets of tracers). According to Latorre et al. [13], the outputs of CTS method were efficient in extracting solutions for the study case.

The methods applied to quantify the uncertainty associated with sediment fingerprinting are commonly based on the Monte Carlo simulation technique. For each tracer from each source groups/sink an estimate of the mean is obtained considering its probability distributions, and thus optimized solutions of the mixing model are obtained repeatedly [6, 32, 34–36]. From this modelling Franks and Rowan [6] developed a model that allowed to determine the confidence intervals on un-mixing model results. In this analysis, the variability of each source and number of samples used influenced the model performance. However, according to [37] it is necessary to consider weightings to take into account of the within-source variability and discriminatory power of individual tracer properties and correction factors (particle size and organic matter content for source type or spatial source category) to calculate the relative contribution of sources. In another study, Laceby and Olley [5] proposed a distribution mixing model, whose approach reincorporates correlations between elemental concentrations and models distributions for source contribution terms for multiple targets (end-members). This study demonstrated how different weightings can affect modelling results. Nevertheless, correlation between tracers are not considered.

In this sense, Clarke [36] and Clarke and Minella [38] proposed a statistical method that quantifies the number of samples required to characterize sources and sediments on estimating the uncertainties of the results. In addition, they added the effect of

collinearity between tracer variables to the analysis. In this approach, the variance-covariance matrix is used to solve the over-determined system; the authors included the correlations between variables (covariance) and the effect of the variance in each tracer variable. They also used the Mahalanobis distance to determine the confidence region associated with the uncertainties, and the model shows how the area of the confidence region varies with a decreasing number of samples available. This is one of the main conceptual differences compared to the conventional model [18], which does not consider multicollinearity among tracer properties and uses the mean concentrations of tracing properties in the available samples. We consider that these assumptions limit the accuracy of the results and prevent a broader uncertainty analysis.

In general, numerical processing computer programs are necessary to calculate the contributions of each source to target sediments. In this sense, sharing algorithms and codes that can be tested, modified, and redistributed among researchers and the general public is of utmost importance to develop new techniques more quickly and efficiently and improve existing models [2, 17]. Thus, free, open-source, and multi-platform programming languages that are easy to learn and use are good options for developing numerical programs that will be available to other researchers, especially for researchers who have limited programming skills. In this sense, the the growth of the number of sediment source identification techniques relies on the improvement of algorithms coded in computer languages that translate mathematical and statistical modeling into an operational, standardized, and open-source code/language accessible to the entire scientific community.

Despite the challenges raised by studies of the last decades, the sediment fingerprinting technique has proven to be efficient although some further improvement potential remains possible among the scientific community [15–17, 39]. Many efforts have been directed at creating protocols to standardize monitoring and modeling techniques, facilitating information and knowledge sharing. Accordingly, the goal of the current research is to present and make available a computational structure (module) to solve over-determined systems applied to sediment tracing (or similar areas) and systematize the uncertainty analysis and optimization of the number of samples, which remained under-investigated in the literature during the last years.

In this context, the models proposed by Walling and Woodward [18] and Clarke and Minella [38] will be used to calculate the contributions of each potential source to suspended sediments collected in rivers. The model outputs will be validated using a dataset available from the experimental Arvorezinha catchment, in southern Brazil. The alternative approach proposed in the current research makes it possible to 1) calculate the relative contribution of each source to the composition of the suspended sediment, allowing us to evaluate the effects of reducing the number of samples and the associated uncertainties; 2) consider the possible correlations that inherently exist among the different geochemical variables that compose the set of tracers as determined from the analysis/use of the variance-covariance matrix in the GLS method; and 3) calculate the uncertainty variations associated with a change in the number of samples defining the confidence region of the feasible solutions using the Mahalanobis distance. The Python programming language was chosen to implement the algorithm because it is an open computational language with numerous modules and

libraries available for numerical computation, statistics, data processing, and visualization. Libraries such as Pandas, Numpy, Scipy, Tensorflow, and many others are well-established among data scientists and they allow for rapid prototyping and experimentation. To make this tool available to the entire scientific community, a repository on the GitHub platform was created to make the Python function module available for data analysis, visualization, and mathematical routines applied to sediment tracing modeling.

## 2 Material and methods

The theoretical model applied to the "sediment fingerprinting approach" will be presented in this section. This approach uses tracers analysed in both soil (i.e., potential sources) and eroded sediments in multivariate statistical analyses and mathematical models to optimize over-determined systems. Considering that the suspended sediments transiting in rivers originate from a mixture of eroded soil from different sources, the method is based on the principle that the suspended sediment conserves some bio-physico-chemical characteristics of the sources from which it originates, which is reflected in its composition. These characteristics that remain "conservative" and discriminant are considered potential tracer properties; we can test their efficiency in distinguishing sediment sources and determining the contribution of each source to a given sample collected in the catchment outlet. Nevertheless, various assumptions underlying this approach must be verified and they strongly depend on the number of samples used to characterize a given tracer (and its spatial variability) in the sources and eroded sediments.

The techniques employed in the fingerprinting approach usually follow three distinct stages of statistical analysis: 1) range test, 2) discrimination and 3) classification. The first step refers to determining which geochemical characteristics have the potential to be selected as tracers among the sources; this step is known as discrimination analysis. The second step consists of classifying the eroded sediment samples into the $n$-dimensional space defined by the source tracer properties. In this step, the relative contribution of each of the sources to the suspended sediment composition is calculated by solving a system of over-determined linear equations.

The least squares estimation method based on the well-known Gauss-Markov theory has played an essential role in estimating the unknown parameters in linear regression models [21]. The main assumptions of the technique are that the errors are assumed to be independent and normally-distributed random quantities with zero mean and common variance $\sigma^2$. In this case, its least squares estimator is the best unbiased linear estimator of any linear combination of the observations. There are many ways to define the "best" solution, and one choice is to minimize the sum of squares of the residuals, where the "best" solution means that the least squares estimators of its parameters have minimum variance. When there is a certain degree of correlation between the residuals in a regression model, the ordinary least squares (OLS) and weighted least squares (WLS) may be statistically inefficient or even rely on misleading inferences, in which case it should be preferred to apply the GLS technique to estimate the unknown parameters in a regression model [23–25]. In addition, the

least squares minimization technique allows to estimate regression parameters under constraints [19, 20].

## 2.1 Review of the classical sediment fingerprinting model

The model developed byYu and Oldfield [40] provides a mathematical formulation to obtain the relative contribution of each of the sources to the suspended sediment composition. This model is based on the mass balance of the $m$ tracers in the different potential sources $g$. Therefore, the concentrations of $m$ tracers measured in each suspended sample sediment are written as the linear combination of the concentrations of $m$ tracers measured in each of the different potential sources. The system of equations resulting from this expression is given by

$$y_i = \sum_{s=1}^{g} x_{si} P_s \tag{1}$$

where $i = 1, ..., m$ the number of tracers. The quantity $x_{si}$ is the mean concentration of the $i-$th tracer in the $s-$th sediment source and is estimated from samples collected in each sediment source; $P_s$ $(s = 1, 2, ..., g)$ the proportions of sediment supplied by the $g$ sources and $y_i$ is the concentration of the $i-$th tracer in the suspended sediment $(i = 1, ..., m)$.

Additional equations are assumed to ensure the feasibility of the solutions, since $P_1 + P_2 + ... + P_g = 1$, with all the $P_s$ commonly defined as non-negative and lower than unity, that is $0 \leq P_s \leq 1$, $s = 1, ..., g$. Furthermore, the samples in each potential source and in the suspended sediment are assumed to be statistically independent.

### 2.1.1 Constrained least-squares optimization

Regression analysis is one of the most widely used statistical tools because it provides simple methods for establishing a functional relationship among variables [19–21]. Constrained optimization procedures to estimate the proportions $P_s$ by least-squares objective function (OLS) were first used by Walling and Woodward [18]. The authors estimated the $P_s$ by minimization of

$$\sum_{i=1}^{m} \left\{ \left( y_i - \left( \sum_{s=1}^{g} x_{si} P_s \right) \right) / y_i \right\}^2 \tag{2}$$

Subject to the constraints

$$0 \leq P_s \leq 1 \tag{3}$$

and

$$\sum_{s=1}^{g} P_s = 1 \tag{4}$$

The division by $y_i$ $(i = 1, \cdots, m$ the number of tracers) provides a type of scaling to data for different tracers, which may differ greatly in variability.

It should be noted that, provided the number of source areas $(g)$ is not greater than the number of tracers $(m)$, criterion (2) could in principle be minimized by the least

squares, even if there was only one sample from each sediment source and only one sample of suspended sediment. In this case, an objective function is used to minimize Eq. (2).

This model has provided the main strategy applied by researchers to estimate the relative contribution of each of the sources to the sediment composition. However, the approach is associated with some limitations regarding the influence of the number of samples used to conduct it, since it considers the mean value of the sample properties and it does not take into account collinearity between the variables used. Both limitations have a strong impact on the quality of the results and prevent an uncertainty analysis.

Basead on the model proposed by Walling and Woodward [18], Clarke [36] and Clarke and Minella [38] applied the OLS to the original data (OLS_Clarke model) and introduced the constraints on solving the system, following the steps below:

**1º)** The division of the terms of the Equation 1 by $y_i$:

$$1 = \sum_{s=1}^{g} \left(\frac{x_{si}}{y_i}\right) P_s = \left(\frac{x_{1i}}{y_i}\right) P_1 + \left(\frac{x_{2i}}{y_i}\right) P_2 + ... + \left(\frac{x_{gi}}{y_i}\right) P_g$$

So, the equation is rewritten:

$$1 = \sum_{s=1}^{g} a_{si} P_s \tag{5}$$

where $a_{si} = \dfrac{x_{si}}{y_i}$.

Nominating $\mathbf{W}$ the matrix generated by $\mathbf{a_s}$, $s = 1 \cdots g$ the column vectors ($i = 1, \cdots, m$ the number of tracers), $\mathbf{p}$ column vectors of $P_s$ and $\mathbf{l}$ column vectors of 1, then the underdetermined system of linear equations (Eq. 5) is written matricial form

$$\mathbf{Wp} = \mathbf{l} \tag{6}$$

**2º)** An ordinary least squares method (OLSE) is equivalent to solving the minimization problem

$$\min_{p} \left(\mathbf{Wp} - \mathbf{l}\right)^T \left(\mathbf{Wp} - \mathbf{l}\right), \tag{7}$$

The associated normal system to Eq. 6 is given by

$$\mathbf{W^T W p} = \mathbf{W^T l}, \tag{8}$$

**3º)** Adding the constraints $P_1 + P_2 + ... + P_g = 1$ to the normal system of equations associated with Eq. (8) results as Eq. (9):

$$\mathbf{AP} = \mathbf{Z} \tag{9}$$

where $\mathbf{A}$ is $(g+1)\times(g+1)$ matrix, $\mathbf{P}$ is $(g+1)\times 1$ vector and $\mathbf{Z}$ is $(g+1)\times 1$ vector are given, respectively, by

$$\mathbf{A} = \begin{pmatrix} \langle \mathbf{a_1}, \mathbf{a_1} \rangle & \langle \mathbf{a_1}, \mathbf{a_2} \rangle & \cdots & \langle \mathbf{a_1}, \mathbf{a_g} \rangle & 1 \\ \langle \mathbf{a_2}, \mathbf{a_1} \rangle & \langle \mathbf{a_2}, \mathbf{a_2} \rangle & \cdots & \langle \mathbf{a_2}, \mathbf{a_g} \rangle & 1 \\ \vdots & \ddots & & \vdots & \\ \langle \mathbf{a_g}, \mathbf{a_1} \rangle & \langle \mathbf{a_g}, \mathbf{a_2} \rangle & \cdots & \langle \mathbf{a_g}, \mathbf{a_g} \rangle & 1 \\ 1 & 1 & \cdots & 1 & 0 \end{pmatrix} \tag{10}$$

$$\mathbf{P} = \begin{pmatrix} P_1 \\ P_2 \\ \vdots \\ P_g \\ 1 \end{pmatrix} \tag{11}$$

$$\mathbf{Z} = \begin{pmatrix} \langle \mathbf{a_1}, \mathbf{1} \rangle \\ \langle \mathbf{a_2}, \mathbf{1} \rangle \\ \vdots \\ \langle \mathbf{a_g}, \mathbf{1} \rangle \\ 1 \end{pmatrix} \tag{12}$$

where $\mathbf{a_s}$ for $s = 1 \cdots g$ are the dimensionless column vectors of $\mathbf{W}$ matrix.

**4º)** Then, Equation 9 must be solved.

$$\widehat{\mathbf{P}} = \mathbf{A^{-1}Z} \tag{13}$$

A solution of Equation 9 will automatically satisfy the constraint $P_1+P_2+\cdots+P_g = 1$, although it may not satisfy the inequality constraints $0 < P_1, P_2, \cdots, P_g < 1$. In this case, the solution must be tested in order to satisfy the inequality to obtain all the feasible solutions.

In this approach, the authors were able to analyze the influence of reducing the number of samples used in the uncertainty analysis. However, the model did not take into account the collinearity between the potential tracing variables.

### 2.1.2 The GLS_Clarke model

The method developed by Clarke and Minella [38] presents a way to calculate the uncertainty of the sources apportioment when the number of samples of sources and/or target varies. Besides, the multicollinearity between potential tracing properties is incorporated from the variance-covariance matrix (GLS). For the GLS_Clarke model, the authors suggest the following procedure:

**1º)** The dimensional underdetermined system

$$y_i = \sum_{s=1}^{g} x_{si} P_s \tag{14}$$

is written in matrix by naming $\mathbf{X}$ as the matrix generated by $\mathbf{x_s}$, $s = 1 \cdots g$ the column vectors ($m$-dimensional vectors of the tracers), $\mathbf{p}$ column vectors of $P_s$ and $\mathbf{Y}$ column vectors of concentration in suspended sediment. Then, the under-determined system of linear equations (Eq. 14) is given by

$$\mathbf{Xp} = \mathbf{Y} \tag{15}$$

**2º)** A generated least squares method (GLSE) is equivalent to solving the minimization problem

$$\min_{p} \left(\mathbf{Xp} - \mathbf{Y}\right)^T \mathbf{S^{-1}} \left(\mathbf{Xp} - \mathbf{Y}\right), \tag{16}$$

where $\mathbf{S}$ is the $(m \times m)$ variance–covariance matrix of the "dependent" variable $y_i$ , $i = 1 \cdots m$.

The associated normal system to Eq. 15, which is known as the Aitken equations ([22]) is given by

$$\mathbf{X^T S^{-1} Xp} = \mathbf{X^T S^{-1} Y}, \tag{17}$$

**3º)** Adding the constraints $P_1 + P_2 + ... + P_g = 1$ to the normal system of equations associated Eq. (17) results into:

$$\mathbf{BP} = \mathbf{V} \tag{18}$$

where $\mathbf{B}$ is $(g+1) \times (g+1)$ matrix, $\mathbf{P}$ is $(g+1) \times 1$ vector (Eq. 11) and $\mathbf{V}$ is $(g+1) \times 1$ vector are given, respectively, by

$$\mathbf{B} = \begin{pmatrix} \langle \mathbf{x_1}, \mathbf{b_1} \rangle & \langle \mathbf{x_1}, \mathbf{b_2} \rangle & \cdots & \langle \mathbf{x_1}, \mathbf{b_g} \rangle & 1 \\ \langle \mathbf{x_2}, \mathbf{b_1} \rangle & \langle \mathbf{x_2}, \mathbf{b_2} \rangle & \cdots & \langle \mathbf{x_2}, \mathbf{b_g} \rangle & 1 \\ \vdots & \ddots & & \vdots & \\ \langle \mathbf{x_g}, \mathbf{b_1} \rangle & \langle \mathbf{x_g}, \mathbf{b_2} \rangle & \cdots & \langle \mathbf{x_g}, \mathbf{b_g} \rangle & 1 \\ 1 & 1 & \cdots & 1 & 0 \end{pmatrix} \tag{19}$$

$$\mathbf{V} = \begin{pmatrix} \langle \mathbf{x_1}, \mathbf{y} \rangle \\ \langle \mathbf{x_2}, \mathbf{y} \rangle \\ \vdots \\ \langle \mathbf{x_g}, \mathbf{y} \rangle \\ 1 \end{pmatrix} \tag{20}$$

where $\mathbf{b_s}$ and $\mathbf{y}$ for $s = 1 \cdots g$ are the column vectors of the product of matrices $\mathbf{S^{-1}}$ by $\mathbf{X}$ and $\mathbf{Y}$, respectively. In that way, entries of vectors are given by

$$b_{jk} = \sum_{l=1}^{m} s_{jl} \alpha_{lk} \tag{21}$$

and

$$y_j = \sum_{l=1}^{m} s_{jl} y_l \tag{22}$$

for $j = 1 \cdots m$ and $k = 1 \cdots g$.

**4º)** Equation 18 can then be solved:

$$\widehat{\mathbf{P}} = \mathbf{B^{-1}V} \tag{23}$$

where $\widehat{\mathbf{P}}$ is the vector solution to the sediment source contributions.

### 2.1.3 Mahalanobis distance

The Mahalanobis distance is a multivariate distance metric that measures the distance between a point (vector) and a distribution. This distance differs from the Euclidean distance because it is calculated using the inverse of the variance-covariance matrix of the dataset [41, 42]. It is a useful metric since it measures the distances taking into account the correlation between the variables, even if the data are not normalized values (different scale). The Mahalanobis distance was used to determine the confidence region and estimate the associated uncertainties. Let $n^*$ be the number of feasible solutions of the systems (13 or 23), thereby for each $j = 1 \cdots n^*$ there is the solution vector $\mathbf{P}_j = [P_1\ P_2\ \cdots\ P_{g-1}]^T$, and a vector of the averages of the feasible solutions $\overline{\mathbf{P^*}} = [\overline{P_1}\ \overline{P_2} \cdots \overline{P_{g-1}}]^T$, both of dimension $(g-1) \times 1$. Thus, the Mahalanobis distance is defined by:

$$d_j{}^2 = (\mathbf{P}_j - \overline{\mathbf{P^*}})^T \Sigma^{-1} (\mathbf{P}_j - \overline{\mathbf{P^*}}) \tag{24}$$

where $\Sigma$ is the variance-covariance matrix $(g-1) \times (g-1)$ of the feasible solutions $n^*$.

## 2.2 Case study: Arvorezinha experimental catchment

The dataset used for the evaluations of the mathematical models (OLS_Clarke model and GLS_Clarke model) explored in the Sections Sections 2.1.1 and 2.1.2 was taken from a sediment tracing study performed in the Arvorezinha experimental catchment between 2002 and 2006 (Figure 4). This catchment is located in southern Brazil ($28^o52'$ S and $52^o05'$ O), it covers a surface area of 1.19 $km^2$, and is located on the edge of the Brazilian southern plateau. In its upper part, the topography is gently rolling, and in its lower parts, it is characterized by shorter and steeper slopes. Volcanic rocks and shallow and fragile soils (Entisols and Inceptisols) characterize the geology and the pedology of the catchment. The climate is subtropical super-humid meso-thermic (i.e., Cfb). The mean annual precipitation is 1605 mm over 50-yr period, evenly distributed throughout the year. Land use is mainly agricultural, with much of the land used for tobacco cultivation. Soil erosion is the main soil degradation process thereby generating high sediment yields Minella et al. [43].

Insert Figure 1 here

This experimental catchment is used to investigate hydrological and erosion processes at the catchment scale in this region of South America. The primary goal of

this research was monitoring liquid and solid discharges during significant rainfall-flow events. This monitoring data is then used to improve water and sediment transfer modeling techniques, and to identify and quantify sediment source contributions through the implementation of sediment fingerprinting techniques [3, 44]. In the study of Minella et al. [43], three potential sediment sources were considered, and their relative contributions to 24 suspended sediment samples collected during 20 significant rainfall-flow events were quantified. The sediment sources evaluated were channel banks ($CB$), unpaved roads ($UR$) and crop fields ($CF$). For tracer determination, 9 $CB$ samples, 9 $UR$ samples and 20 $CF$ soil samples were collected across the catchment. The set of 62 samples of suspended sediments and potential sources were characterized for their elemental geochemistry (total concentrations in P, K, Ca, Na, Mg, Cu, Pb, Cr, Co, Zn, Ni, Fe, Mn, and Al). As a first step, the individual analysis of the discriminant ability of each element was performed by means of a range test to evaluate the conservativity and the Kruskal-Wallis test to estimate the discrimination power [7]. Subsequently, the best set of tracer elements was determined based on multivariate or discriminant functional analyses to define the best set of variables for discriminating sources. This analysis maximizes the discrimination between the sources and minimizes the number of variables required. The method is based on minimizing the Wilks' Lambda index ($\lambda$*), a component of the multivariate analysis of variance.

The methods of tracer selection are fundamental in sediment fingerprinting because it may decrease the deleterious effect of multicollinearity. Davis and Sampson [45] reported that redundant variables may weaken the analysis due to the reduction of the degrees of freedom of the errors and may affect the feasibility the inversion operation of the variance-covariance matrix. In addition, they reduce the dimensionality of the problem, significantly improving the ability of the model to find a solution and the associated errors. Seven ($m = 7$) of the 14 chemical elements previously analyzed were selected to compose the optimal data set in this case study (Fe, Mn, Cu, Zn, Ca, K and P are the optimal tracers).

## 2.3 Fingerprinting computational framework applied to Arvorezinha catchment data set

The Python programming language was used to implement the OLS_Clarke and GLS_Clarke models. The samples collected from $g$ sources and suspended sediments, with the respective concentrations of $m$ tracers with $g \leq m$, provide the input data of the algorithm. It consists of a module and external functions used repeatedly for the approximate resolution of the over-determined linear system with $m+1$ equations and $g$ unknowns in the form $y_i = \sum x_{si} P_s$ with $s = 1, ..., g$ and $i = 1, ..., m$ conditioned on $1 = P_1 + P_2 + ... + P_g$. From the solutions obtained, the feasible proportions provide a means to estimate the uncertainty of the results as a function of reducing the number of samples used. The process is repeated 100 times, generating a cloud of feasible solutions in the $P_1 P_2$ plane. A flowchart was developed to provide an overview of the structural organization of the algorithm.

Insert Figure 2 - Flowchart here

Each block in the flowchart has tasks executed by commands and functions described below:

1. Python import library: The import statement Python.
2. Data set: The data representing each $g$ source and suspended sediment should be organized in $g+1$ spreadsheets, where each column represents one of the optimal tracers and each row represents a particular sample. For instance, spreadsheet $CB$ will be a $9 \times 7$ table that corresponds to 9 samples collected in channel banks and 7 optimal tracers analysed in these, spreadsheet $UR$ will then be a $9 \times 7$ table that contains 9 samples collected unpaved roads and 7 optimal tracers, spreadsheet $CF$ will finally correspond to a $20 \times 7$ table that contains 20 samples collected crop fields and 7 optimal tracers, and spreadsheet $Y$ will include a $24 \times 7$ table that associates 24 suspended sediment samples and their 7 optimal tracers. These subsets are saved in four matrices that will then be used in the next operations.
3. Subset random: It randomly chooses, without repetition, the samples from each subset; in this case, we can select $1 \leq nCB \leq 9$, $1 \leq nUR \leq 9$, $1 \leq nCF \leq 20$ and $1 \leq nY \leq 24$. Hence, the total number of possible combinations involving the samples from each set for the coefficients of Equation (1), is equal to $C_n = nY \times nCB \times nUR \times nCF$, which will be reduced as the number of samples used for analysis decreases. For example, we can choose the sequence $\{Y_n\}_{n=1}^{6} = \{24, 20, 16, 12, 8, 4, 2\}$ as the number of suspended sediment samples $Y$ and keep the number of the other subsets fixed, then the sequence of possible combinations is given by $\{C_n\}_{n=1}^{6} = \{38880, 25920, 19440, 12960, 6480, 3240\}$. Each selected $CB$, $UR$ and $CF$ sample will correspond to a column of the source matrix ($W$ ou $X$) and $Y$ to a column vector of the suspended sediment.
4. Models: For each of the drawings, the external functions OLS_Clarke and GLS_Clarke described in Sections 2.1.1 and 2.1.2 are called.
5. Cloud Proportions: It calls a third external function to compute the proportions $P_1, P_2, ..., P_g$ of the $g$ sources. Among the solutions obtained in the OLS_Clarke and GLS_Clarke procedures, the algorithm retains those that are feasible (i.e., such that $0 < P_1, P_2, P_3 < 1$). The process is then repeated 100 times, generating a cloud of feasible solutions in the $P_1P_2$ plane.
6. Confidence regions: The confidence interval region (95%) is calculated from the set of feasible solutions $n^*$ for each $nY$ or $nCF$ using the procedure described in Section 2.1.3. Lastly, a function generates the graphical visualization of this region.
7. Outputs: The calculation of the mean of the areas of the confidence region, standard deviation, number of combinations, number of feasible solutions, mean of the values of $P_1, P_2, P_3$, and the coefficient of variation.

The Fingerprinting repository on the GitHub platform was created to provide a module of functions in Python for data analysis, visualization, and mathematical routines applied to sediment tracing modeling. The model proposed by [38] is implemented and can be modified and applied to different databases. The module consists of functions in Python and has the libraries Numpy [46], Scipy [47], and MatplotLib [48] as the dependencies. A Jupyter Notebook file provides examples of using some functions and reproducing the results published in this work. From

this example, we can explore other data sets, modify input parameters in the functions, operationalize models, and create graphs. The repository can be accessed at https://github.com/tiagoburiol/Fingerprinting.

# 3 Results and discussion

With the algorithm implemented in Python, it is possible to execute the sequence of instructions that operationalize the OLS_Clarke and GLS_Clarke models and the uncertainty calculation. The simulations allow the user to compare the results generated by both methods to solve over-determined systems, select solutions utilizing the constraints, calculate the contributions of each source to the sediments, and statistical parameters such as the confidence interval region, the standard deviation, and the coefficient of variation. The user can also indicate the number of samples in the different groups of samples (target sediments or soil sources) to analyze the effect of reducing their number on the calculations of the $P_s$ values and the impact of this reduction on the uncertainty associated with the output results.

From the simulation results, we quantified the influence of collinearity on the results of the source ratios. In the tables and figures below, the results for the OLS_Clarke model correspond to the application of the model without considering the effect of collinearity, and the results for the GLS_Clarke model, which indicate the application of the model when considering the effect of collinearity. The comparison between the two models is obtained simultaneously, that is, using the same subset of randomly drawn samples.

In addition to the effect of collinearity, we also present the analysis of the increased uncertainties associated with a reduction in the number of samples from both the suspended sediment samples ($nY$) and the crop fields samples ($nCF$). The source $CF$ was choose the higher number of data. The two simulation to evaluate the sample number reduction were performed using the OLS_Clarke and GLS_Clarke models under constraints described in Sections 2.1.1 and 2.1.2. The values of $P_1$, $P_2$,and $P_3$, which correspond to the contributions from the $CB$, $UR$, and $CF$ sources, are identified by: a) Simulation 1: reduction of suspended sediment samples $nY = \{24, 20, 16, 12, 8, 4, 2\}$ (Table 1) and b) Simulation 2: reduction of the number of crop samples $nCF = \{20, 16, 12, 8, 4, 2\}$ (Table 2).

Insert TABLE 1 here

Insert TABLE 2 here

The first result is that the inclusion of the variance-covariance matrix, and consequently the effect of collinearity, affects the results of both simulations. The significant difference in the control parameters (mean area, $\sigma$ and $n^*$) and in the mean values of $P_1$, $P_2$, and $P_3$ can also be observed.

Although both models indicate the $CF$ source ($\overline{P_3}$) as the main source of sediment, the OLS_Clarke model suggests that the contribution from the CB source ($\overline{P_1}$) is greater than that from the $UR$ source ($\overline{P_2}$). Nonetheless, the GLS_Clarke model

indicates that the $UR$ source ($\overline{P_2}$) supplies more sediment than the $CB$ source ($\overline{P_1}$), corroborating field evidence and observations. The catchment is characterised by a dense road network, where many sections are inadequately allocated given the local topography and without the implementation of runoff control practices. According to a study by [49] in this catchment, unpaved roads are severely affected by water erosion and contribute significantly to sediment yield.

These results suggest that it is important to take into consideration the possible correlations between the chemical variables that comprise the tracer set as the GLS method performed. In this model, the values of $n^*$ are lower, and the value of $\sigma$ decreases significantly, although the average area of the confidence region is larger.

The comparison between both methods demonstrates the interest of extracting the effect of collinearity existing in the set of tracers even if these have been appropriately selected in the discrimination phase by the range test, KruskalWallis test, and minimization of Wilk's lambda. According to Johnson et al. [50] the variance-covariance matrix allows to extract the effects of correlation among the variables in the solution of the systems as those presented in this work. De Maesschalck et al. [41] through the Mahalanobis distance, also demonstrated the applicability of this approach in reducing the uncertainties associated with collinearity.

Figure 3 shows the point distributions that express the feasible solutions in the 95% confidence interval region. In the GLS_Clarke model, the point distributions show a higher "density" in the region near the mean value than those obtained from the OLS_Clarke model.

Insert Figure 3 here

However, it is important to note that the possibility of considering collinearity among the tracer variables does not diminish the importance of previous discrimination analyses in defining the best set of tracers. Including the variance-covariance matrix of the previously selected variables considers the existing correlation between the chosen variables, even if it remains limited. It is known that sediment production in catchments can result from different processes (diffuse or concentrated, agricultural or fluvial, superficial or deep erosion) in which different variables will reflect the set of operating processes. Therefore, we consider it fundamental to analyse a significant amount of tracers that maximize the discriminating capacity, which will offer the physico-chemical basis to differentiate the sources.

Even after the selection analysis of the set of variables in the minimization of the Wilks' Lambda index ($\lambda*$), a certain degree of collinearity may influence the final result. Moreover, to exclude the variables (tracers) that present some degree of collinearity would be neglecting part of the useful existing variability. This fact can be verified in the higher uncertainties by the simulations of the OLS_Clarke model that does not consider the variance-covariance matrix in the solution of the over-determined equation systems.

In Simulation 1 and for both models, the number of feasible solutions and the mean area of the confidence region decreased as the number of suspended sediment samples considered decreased, corroborating the result of Clarke and Minella [38].

The distribution of points expressing the feasible solutions within the confidence interval region (95%) for the cases $nY = 12$ and $nY = 4$ are illustrated in Figures 4 and 5, respectively. We can observe the lower number of feasible solutions. Besides this, and, as in Figure 3, the points are more concentrated around the $P_1$ and $P_2$ mean values in the GLS_Clarke model.

Insert FIGURE 4 here

Insert FIGURE 5 here

Figure 6 shows the coefficients of variation ($CV$) obtained with the reduced number of sediment samples for each model (OLS_Clarke and GLS_Clarke). In these plots, we can observe the impact of reducing the number of eroded sediment samples collected in the river when defining the $P_s$ values. For all sample quantities, the standard deviation is lower with the GLS_Clarke model. Additionally, the difference in the $CV$ rates of change in the OLS_Clarke model is much higher in a few samples (e.g., in the range below ten samples).

Insert FIGURE 6 here

Figures 7, Fig. 8, and Fig.9 show the regions of the confidence interval (95%) with the reduction in the number of samples from $nCF = 20$, $nCF = 12$, and $nCF = 4$, respectively. Notably, the variance-covariance matrix in the GLS_Clarke model comes from the $CF$ source data. Therefore, the point distributions in Figures 6 ($b$) and 7 ($b$) differ even though both simulations use the total number of existing samples (ie $nY = 24$ and $nCF = 20$).

Insert FIGURE 7 here

Insert FIGURE 8 here

Insert FIGURE 9 here

As for the previous scenario simulation, the number of feasible solutions decreases as the number of suspended sediment samples considered decreases, as does the average area of the confidence region. The cases simulated by both models have mean proportions of $\overline{P_1}$ greater than the mean proportions of $\overline{P_2}$, indicating that the relative contribution to the $CB$ source is greater than that of $UR$ in the suspended sediment composition ($\overline{P_1} > \overline{P_2}$). This result differs from that of Simulation 1, in which the GLS_Clarke model showed similar solutions as the OLS_Clarke model with $\overline{P_2} < \overline{P_1}$. This indicates that the GLS method responds directly to the choice of inclusion of the variance-covariance matrix. Figure 10 presents the $CV$ values associated with the reduction in the number of crop fields samples considered, and Table 2 lists the mean values of the proportions obtained in such a situation.

Insert FIGURE 10 here

The OLS_Clarke model showed higher $CV$ values than the GLS_Clarke model when analysing the uncertainties associated with the results obtained with a reduced number of samples, indicating, once again, that the uncertainties decrease as the effect of collinearity among the variables is considered in the system over-determined by the presence of the variance-covariance matrix.

In all simulated cases, the relative contribution of $\overline{P_3}$ is greater, indicating that the crop fields ($CF$) source provides the main sediment source in the investigated catchment. This result corroborates those of other previous studies conducted by Minella et al. [44] in the catchment and all documented field observations.

# 4 Conclusions

This study presented an updated and open access computational framework for solving over-determined systems applied to sediment tracing based on the models previously proposed by Clarke and Minella [38]. The algorithm enabled us to analyze and compare the use of different metrics, optimize the procedure for calculating the area that expresses the degree of uncertainty associated with the number of samples taken into account to document source and target sediment properties, and consider the possible correlations that naturally exist between the different variables that compose the set of tracers used for sediment tracing.

Our findings confirmed the assumption of increased uncertainty as the number of samples considered decreases either for the potential sources or the target sediment samples. Moreover, considering the variance-covariance matrix to find the solution of the over-determined system allowed us to consider the deleterious effects of collinearity between tracers in sediment tracing studies. The implemented algorithm allowed us to compare the two modeling strategies and simulate multiple scenarios of sample number reduction.

The Phyton language enables the use of an easy-to-manipulate code, and will facilitate the shared implementation of the model and the inclusions of model modifications that may be suggested by the research community to keep improving and standardizing the sediment tracing protocol. With this tool, new perspectives are opened to to provide a helping decision tool to define the number of samples required to characterise the potential sources and/or sediment based on the uncertainty analysis of the set of samples available, which is fundamental for the advancement of research in environmental monitoring and modeling, as well as for the effective management of soil and water resources at the catchment scale.

Considering that methodological development occurs through the contribution of professionals from different areas of knowledge, exchanging ideas and the efficient use of computing resources promotes the standardization and accuracy of techniques. In this context, the sharing of the algorithm implemented in Python provides speed and transparency in scientific development and increases access to new methods related to fingerprinting, providing researchers with an algorithm that enables them to evaluate

uncertainties in reducing the number of samples and the influence of the collinearity of the set of tracers, which can also be adapted to other areas of knowledge.

# Declarations

## Ethical Approval

Not applicable

## Consent to Participate

Not applicable

## Consent to Publish

The authors hereby consents to publication of the manuscript in the Computational and Applied Mathematics journal.

## Authors Contributions

Conceptualization: Lidiane Buligon and Jean Paolo Gomes Minella; Methodology: Jean Paolo Gomes Minella, Tiago Martinuzzi Buriol, and Lidiane Buligon; Formal analysis and investigation: Jean Paolo Gomes Minella, Tiago Martinuzzi Buriol, and Lidiane Buligon; Writing - original draft preparation: Lidiane Buligon, Jean Paolo Gomes Minella, and Tiago Martinuzzi Buriol; Writing - review and editing: Olivier Evrard; Funding acquisition: Lidiane Buligon, Jean Paolo Gomes Minella, and Olivier Evrard.

## Funding

CNPq (National Council for Scientific and Technological Development) Process 200008/2023-4. CAPES (Coordination for the Improvement of Higher Education Personnel) Process 88887.310201/2018-00. CEA (Atomic Energy Commission).

## Competing Interests

The authors declare that they have no conflict of interest that are relevant to the content of this article.

## Availability of data and materials

Data will be made available on reasonable request.

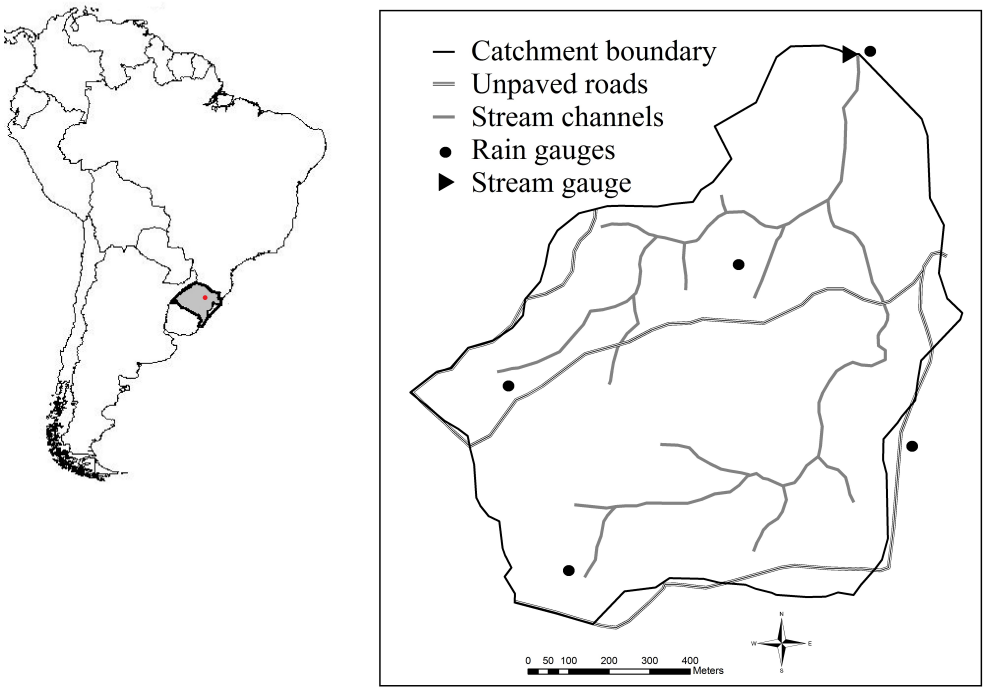



**Figure 1** Arvorezinha experimental catchment: Adapted from [3].

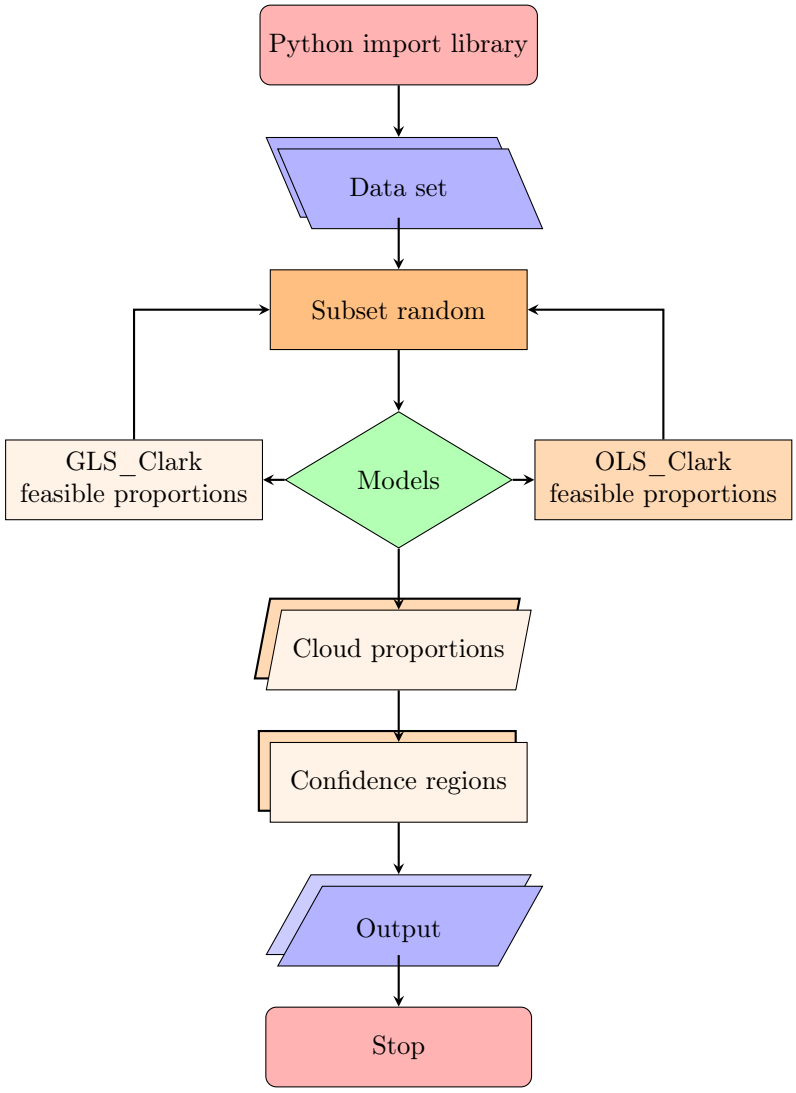
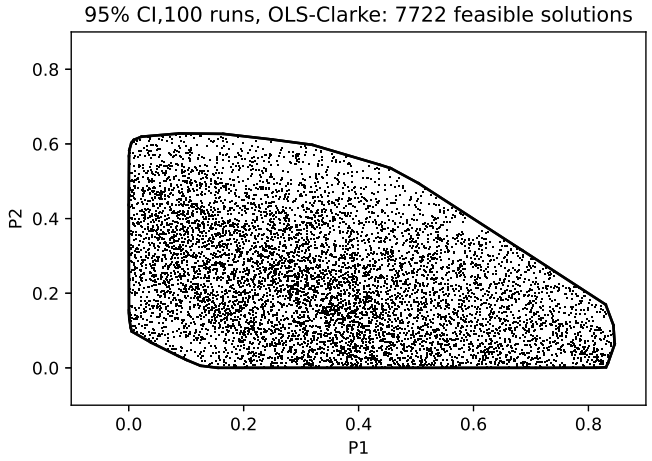
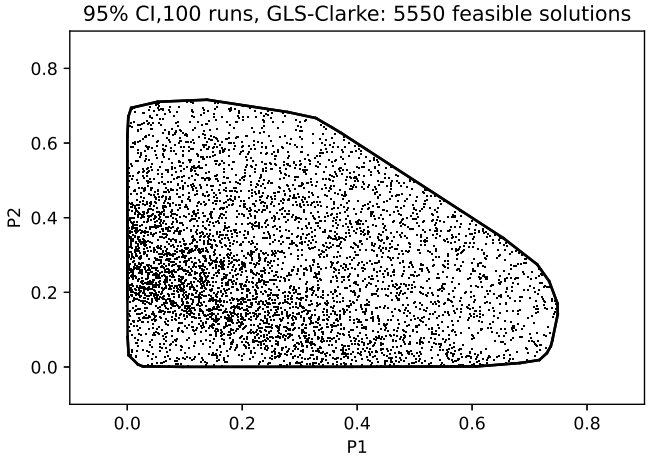



**Figure 2** Flowchart of the structural organization of the algorithm.

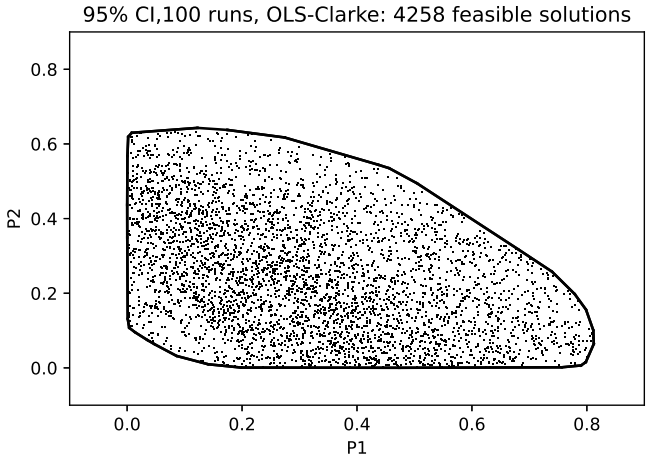



(a) OLS_Clarke

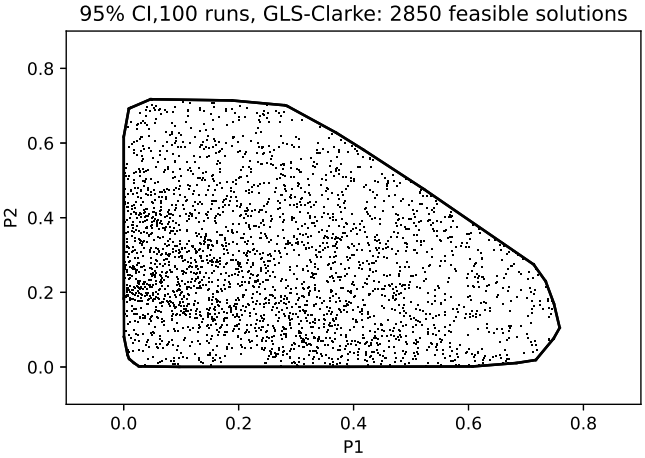



(b) GLS_Clarke

**Figure 3** The feasible solutions in the 95% confidence region. The mean of proportions $(\overline{P_1}, \overline{P_2}) = (0.341, 0.238)$ and $(\overline{P_1}, \overline{P_2}) = (0.259, 0.276)$ obtained by *a*) OLS_Clarke model and *b*) GLS_Clarke model, respectively. Considering $nY = 24$ over 100 repetitions.

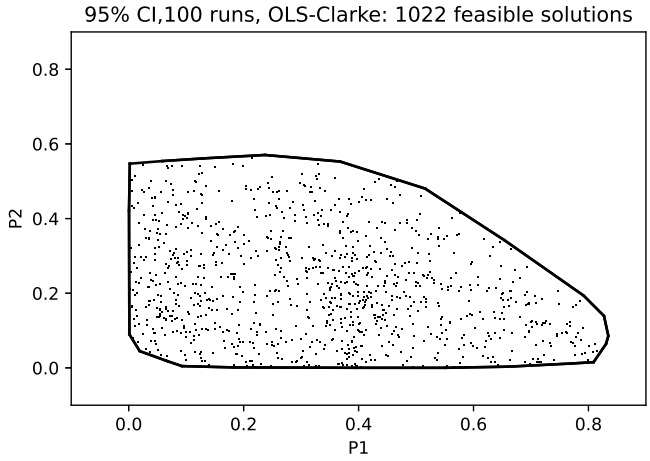



(a) OLS_Clarke

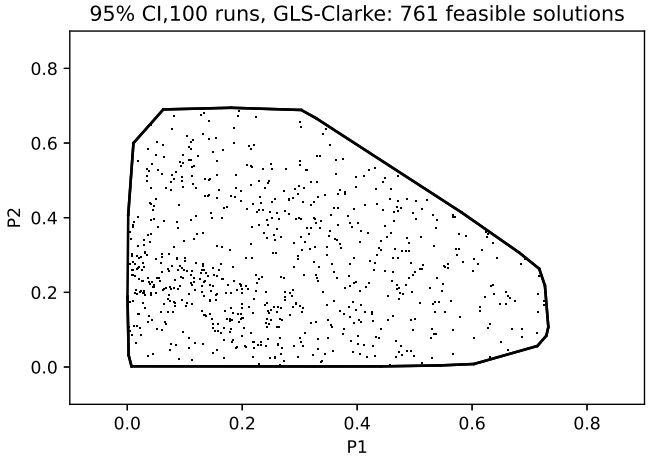



(b) GLS_Clarke

**Figure 4** The feasible solutions in the 95% confidence region. The mean of proportions $(\overline{P_1}, \overline{P_2}) = (0.319, 0.248)$ and $(\overline{P_1}, \overline{P_2}) = (0.267, 0.278)$ obtained by *a*) OLS_Clarke model and *b*) GLS_Clarke model, respectively. Considering $nY = 12$ over 100 repetitions.




(a) OLS_Clarke




(b) GLS_Clarke

**Figure 5** The feasible solutions in the 95% confidence region. The mean of proportions $(\overline{P_1}, \overline{P_2}) = (0.333, 0.219)$ and $(\overline{P_1}, \overline{P_2}) = (0.266, 0.276)$ obtained by *a*) OLS_Clarke model and *b*) GLS_Clarke model, respectively. Considering $nY = 4$ over 100 repetitions.

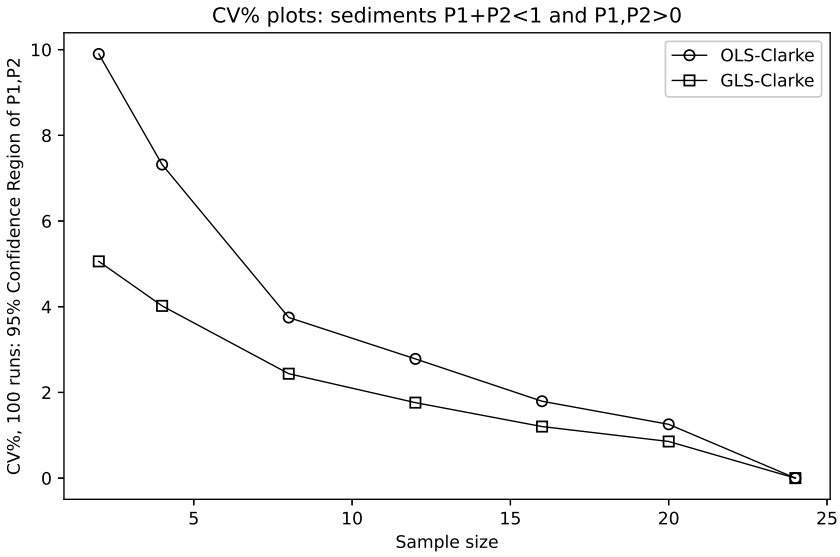
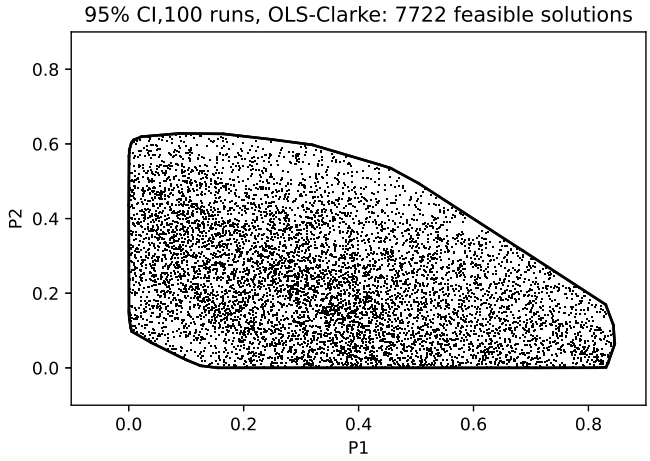
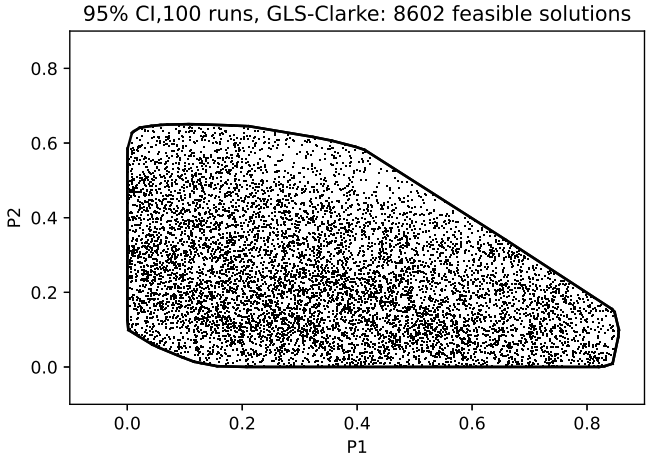



**Figure 6** Relationship between coefficients of variation ($CV$) and the number of suspended sediment samples of the 95% confidence regions for channel banks ($P_1$) and unpaved roads ($P_2$) when samples sizes are reduced by sequence $nY = \{24, 20, 16, 12, 8, 4, 2\}$ using OLS_Clarke and GLS_Clarke models over 100 repetitions.

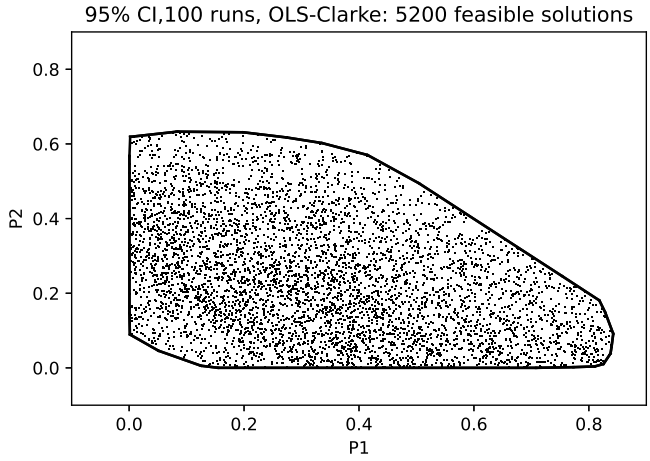



(a) OLS_Clarke

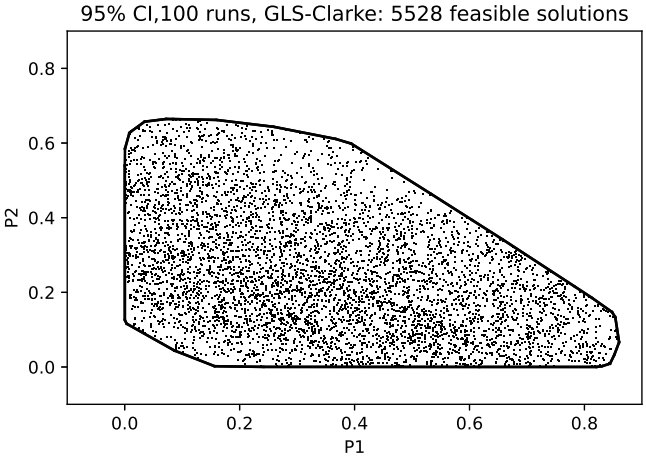



(b) GLS_Clarke

**Figure 7** The feasible solutions in the 95% confidence region. The mean of proportions $(\overline{P_1}, \overline{P_2}) = (0.341, 0.238)$ and $(\overline{P_1}, \overline{P_2}) = (0.341, 0.245)$ obtained by *a*) OLS_Clarke model and *b*) GLS_Clarke model, respectively. Considering $nL = 20$ over 100 repetitions.

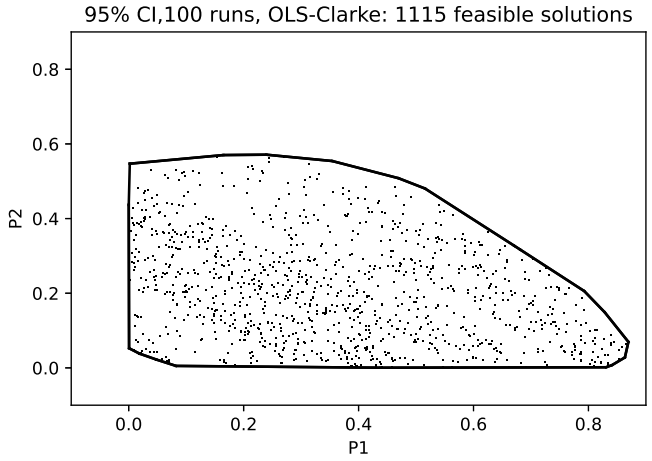



(a) OLS_Clarke

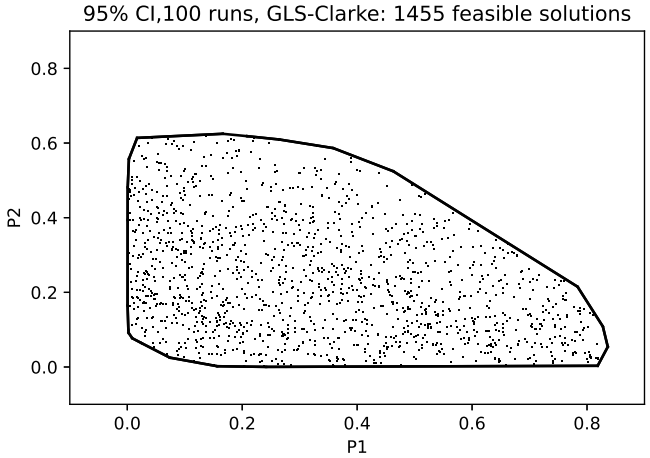



(b) GLS_Clarke

**Figure 8** The feasible solutions in the 95% confidence region. The mean of proportions $(\overline{P_1}, \overline{P_2}) = (0.338, 0.239)$ and $(\overline{P_1}, \overline{P_2}) = (0.348, 0.252)$ obtained by *a*) OLS_Clarke model and *b*) GLS_Clarke model, respectively. Considering $nL = 12$ over 100 repetitions.




(a) OLS_Clarke




(b) GLS_Clarke

**Figure 9** The feasible solutions in the 95% confidence region. The mean of proportions $(\overline{P_1}, \overline{P_2}) = (0.358, 0.208)$ and $(\overline{P_1}, \overline{P_2}) = (0.330, 0.234)$ obtained by *a*) OLS_Clarke model and *b*) GLS_Clarke model, respectively. Considering $nL = 4$ over 100 repetitions.

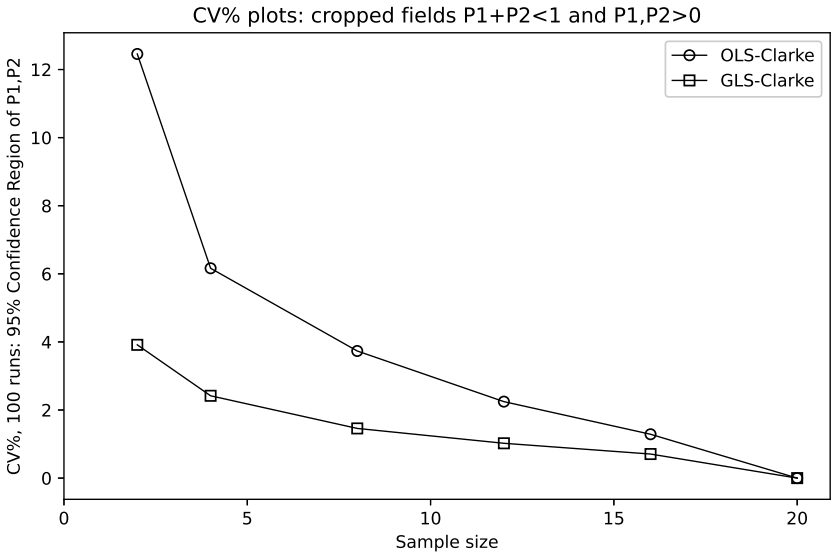



**Figure 10** Relationship between coefficients of variation ($CV$) and the number of suspended sediment samples of the 95% confidence regions for channel banks ($P_1$) and unpaved roads ($P_2$) when samples sizes are reduced by sequence $nCF = \{20, 16, 12, 8, 4, 2\}$ using OLS_Clarke and GLS_Clarke models over 100 repetitions.

**Table 1** Simulation 1 - Considering the reduction of suspended sediment samples ($nY$). *a*) OLS_Clarke model and *b*) GLS_Clarke model.

| *a*) OLS_Clarke model | | | | | | | |
|---|---|---|---|---|---|---|---|
| nY | Mean Area | $\sigma$ | $C_n$ | $n^*$ | $\overline{P_1}$ | $\overline{P_2}$ | $\overline{P_3}$ |
| 24 | 0.403 | 0.000 | 38880 | 7722 | 0.341 | 0.238 | 0.420 |
| 20 | 0.401 | 0.005 | 32400 | 6592 | 0.345 | 0.244 | 0.412 |
| 16 | 0.402 | 0.007 | 25920 | 5163 | 0.316 | 0.243 | 0.440 |
| 12 | 0.399 | 0.011 | 19440 | 4258 | 0.319 | 0.248 | 0.432 |
| 8 | 0.392 | 0.015 | 12960 | 2298 | 0.392 | 0.219 | 0.389 |
| 4 | 0.380 | 0.028 | 6480 | 1022 | 0.333 | 0.219 | 0.448 |
| 2 | 0.366 | 0.036 | 3240 | 779 | 0.325 | 0.318 | 0.356 |
| *b*) GLS_Clarke model | | | | | | | |
| nY | Mean Area | $\sigma$ | $C_n$ | $n^*$ | $\overline{P_1}$ | $\overline{P_2}$ | $\overline{P_3}$ |
| 24 | 0.418 | 0.000 | 38880 | 5550 | 0.259 | 0.276 | 0.466 |
| 20 | 0.418 | 0.004 | 32400 | 4447 | 0.263 | 0.286 | 0.451 |
| 16 | 0.418 | 0.005 | 25920 | 3435 | 0.263 | 0.281 | 0.457 |
| 12 | 0.415 | 0.007 | 19440 | 2850 | 0.267 | 0.278 | 0.456 |
| 8 | 0.413 | 0.010 | 12960 | 2240 | 0.256 | 0.270 | 0.474 |
| 4 | 0.401 | 0.016 | 6480 | 761 | 0.266 | 0.276 | 0.458 |
| 2 | 0.384 | 0.019 | 3240 | 409 | 0.282 | 0.318 | 0.400 |

Number of suspended sediment samples ($nY$); Mean area of the 95% confidence regions; Standard deviations ($\sigma$); Number of possible solutions of the overdetermined systems ($C_n$); Feasible solutions ($n^*$) where $0 < P_1, P_2, P_3 < 1$ and $P_1 + P_2 + P_3 = 1$; Means of proportions contributed by Channel Banks ($\overline{P_1}$), Unpaved Roads ($\overline{P_2}$) and Crop Fields ($\overline{P_3}$) of the 95% confidence regions and over 100 repetitions.

**Table 2** Simulation 2 - Considering the reduction of crop fields samples ($nL$). *a*) OLS_Clarke model and *b*) GLS_Clarke model.

| *a*) OLS_Clarke model | | | | | | | |
|---|---|---|---|---|---|---|---|
| nL | Mean Area | $\sigma$ | $C_n$ | $n^*$ | $\overline{P_1}$ | $\overline{P_2}$ | $\overline{P_3}$ |
| 20 | 0.403 | 0.000 | 38880 | 7722 | 0.341 | 0.238 | 0.420 |
| 16 | 0.401 | 0.005 | 31104 | 5934 | 0.352 | 0.241 | 0.407 |
| 12 | 0.400 | 0.009 | 23328 | 5200 | 0.338 | 0.239 | 0.423 |
| 8 | 0.394 | 0.015 | 15552 | 3154 | 0.350 | 0.232 | 0.418 |
| 4 | 0.386 | 0.024 | 7776 | 1115 | 0.358 | 0.208 | 0.434 |
| 2 | 0.351 | 0.044 | 3888 | 547 | 0.347 | 0.214 | 0.439 |
| *b*) GLS_Clarke model | | | | | | | |
| nL | Mean Area | $\sigma$ | $C_n$ | $n^*$ | $\overline{P_1}$ | $\overline{P_2}$ | $\overline{P_3}$ |
| 20 | 0.414 | 0.000 | 38880 | 8602 | 0.341 | 0.245 | 0.414 |
| 16 | 0.414 | 0.003 | 31104 | 6862 | 0.344 | 0.247 | 0.408 |
| 12 | 0.413 | 0.004 | 23328 | 5528 | 0.348 | 0.252 | 0.400 |
| 8 | 0.410 | 0.006 | 15552 | 3242 | 0.345 | 0.244 | 0.411 |
| 4 | 0.405 | 0.010 | 7776 | 1455 | 0.330 | 0.234 | 0.437 |
| 2 | 0.392 | 0.015 | 3888 | 684 | 0.319 | 0.232 | 0.449 |

Number of crop fields samples ($nCF$); Mean area of the 95% confidence regions; Standard deviations ($\sigma$); Number of possible solutions of the overdetermined systems ($C_n$); Feasible solutions ($n^*$) where $0 < P_1, P_2, P_3 < 1$ and $P_1 + P_2 + P_3 = 1$; Means of proportions contributed by Channel Banks ($\overline{P_1}$), Unpaved Roads ($\overline{P_2}$) and Crop Fields ($\overline{P_3}$) of the 95% confidence regions and over 100 repetitions.

# References


[1] Owens, P.: Soil erosion and sediment dynamics in the anthropocene: a review of human impacts during a period of rapid global environmental change. Journal of Soils and Sediments **20**, 1–29 (2020) https://doi.org/10.1007/s11368-020-02815-9

[2] Collins, A.L., Blackwell, M., Boeckx, P., Chivers, C.-A., Emelko, M., Evrard, O., Foster, I., Gellis, A., Gholami, H., Granger, S., *et al.*: Sediment source fingerprinting: benchmarking recent outputs, remaining challenges and emerging themes. Journal of Soils and Sediments **20**(12), 4160–4193 (2020)

[3] Minella, J.P.G., Walling, D.E., Merten, G.H.: Establishing a sediment budget for a small agricultural catchment in southern brazil, to support the development of effective sediment management strategies. Journal of Hydrology **519**, 2189–2201 (2014) https://doi.org/10.1016/j.jhydrol.2014.10.013

[4] Walling, D.E.: The evolution of sediment source fingerprinting investigations in fluvial systems. Journal of Soils and Sediments **13**(10), 1658–1675 (2013)

[5] Laceby, J.P., Olley, J.: An examination of geochemical modelling approaches to tracing sediment sources incorporating distribution mixing and elemental correlations. Hydrological Processes **29**(6), 1669–1685 (2015) https://doi.org/10.1002/hyp.10287 https://onlinelibrary.wiley.com/doi/pdf/10.1002/hyp.10287

[6] Franks, S., Rowan, J.: Multi-parameter fingerprinting of sediment sources: Uncertainty estimation and tracer selection. Computational methods in water resources - Volume 2 - Computational methods,surface water systems and hydrology, 1067–1074 (2000)

[7] Collins, A., Walling, D.: Selecting fingerprint properties for discriminating potential suspended sediment sources in river basins. Journal of Hydrology **261**, 218–244 (2002)

[8] Cooper, R.J., Krueger, T., Hiscock, K.M., Rawlins, B.G.: Sensitivity of fluvial sediment source apportionment to mixing model assumptions: A bayesian model comparison. Water Resources Research **50**(11), 9031–9047 (2014) https://doi.org/10.1002/2014WR016194 https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1002/2014WR016194

[9] Pulley, S., Collins, A.L.: Tracing catchment fine sediment sources using the new sift (sediment fingerprinting tool) open source software. Science of The Total Environment **635**, 838–858 (2018) https://doi.org/10.1016/j.scitotenv.2018.04.126

[10] Hughes, A.O., Olley, J.M., Croke, J.C., McKergow, L.A.: Sediment source

changes over the last 250 years in a dry-tropical catchment, central queensland, australia. Geomorphology **104**(3), 262–275 (2009) https://doi.org/10.1016/j.geomorph.2008.09.003

[11] Uber, M., Legout, C., Nord, G., Crouzet, C., Demory, F., Poulenard, J.: Comparing alternative tracing measurements and mixing models to fingerprint suspended sediment sources in a mesoscale mediterranean catchment. Journal of Soils and Sediments **19**, 3255–3273 (2019)

[12] Lizaga, I., Latorre, B., Gaspar, L., Navas, A.: Consensus ranking as a method to identify non-conservative and dissenting tracers in fingerprinting studies. Science of the Total Environment **720**, 137537 (2020)

[13] Latorre, B., Lizaga, I., Gaspar, L., Navas, A.: A novel method for analysing consistency and unravelling multiple solutions in sediment fingerprinting. Science of The Total Environment **789**, 147804 (2021)

[14] Batista, P., Laceby, J., Evrard, O.: How to evaluate sediment fingerprinting source apportionments. Journal of Soils and Sediments **22**, 1–14 (2022) https://doi.org/10.1007/s11368-022-03157-4

[15] Collins, A., Pulley, S., Foster, I.D., Gellis, A., Porto, P., Horowitz, A.: Sediment source fingerprinting as an aid to catchment management: a review of the current state of knowledge and a methodological decision-tree for end-users. Journal of Environmental Management **194**, 86–108 (2017)

[16] Laceby, J., Gellis, A., Koiter, A., Blake, W., Evrard, O.: Preface—evaluating the response of critical zone processes to human impacts with sediment source fingerprinting. Journal of Soils and Sediments **19** (2019) https://doi.org/10.1007/s11368-019-02409-0

[17] Evrard, O., Batista, P.V., Company, J., Dabrin, A., Foucher, A., Frankl, A., García-Comendador, J., Huguet, A., Lake, N., Lizaga, I., *et al.*: Improving the design and implementation of sediment fingerprinting studies: Summary and outcomes of the tracing 2021 scientific school. Journal of Soils and Sediments **22**(6), 1648–1661 (2022)

[18] Walling, D., Woodward, J.: Tracing sources of suspended sediment in river basins: A case study of the river culm, devon, uk. Marine and Freshwater Research **46**(1), 327–336 (1995) https://doi.org/10.1071/MF9950327

[19] Chatterjee, S., Hadi, A.S.: Regression Analysis by Example. John Wiley & Sons, Chichester (2013)

[20] Montgomery, D.C., Peck, E.A., Vining, G.G.: Introduction to Linear Regression Analysis. John Wiley & Sons, ??? (2021)

[21] Kariya, T., Kurata, H.: Generalized Least Squares. Wiley Series in Probability and Statistics, p. 312. John Wiley & Sons, Chichester (2004)

[22] Baksalary, J., Puntanen, S.: Weighted-least-squares estimation in the general gauss-markov model. In: Dodge, Y. (ed.) Statistical Data Analysis and Inference, pp. 355–368. Elsevier Science Publishers B.V., ??? (1989)

[23] Ciarlet, P.G., Lions, J.L.: Handbook of Numerical Analysis. Vol. 1: Finite Difference Methods (Part 1) and Solution of Equations in R^n (Part 1). Elsevier Science Publishers, North Holland (1990)

[24] Meyer, C.D.: Matrix Analysis and Applied Linear Algebra. Other Titles in Applied Mathematics, vol. 71, p. 730. Society for Industrial and Applied Mathematics (SIAM), Philadelphia (2000)

[25] Shores, T.S.: Applied Linear Algebra and Matrix Analysis, p. 479. Springer, Berlin (2007). https://doi.org/10.1007/978-3-319-74748-4

[26] Lizaga, I., Latorre, B., Gaspar, L., Navas, A.: FingerPro: an R Package for Tracking the Provenance of Sediment. Water Resources Management: An International Journal, Published for the European Water Resources Association (EWRA) **34**(12), 3879–3894 (2020) https://doi.org/10.1007/s11269-020-02650

[27] Golub, G.H., Van Loan, C.F.: Matrix Computations. Johns Hopkins University Press, Baltimore (2013)

[28] Gentle, J.E.: Matrix algebra. Springer texts in statistics, Springer, New York, NY **10**, 978 (2007)

[29] Lawson, C.L., Hanson, R.J.: Solving least squares problems. In: Classics in Applied Mathematics (1976)

[30] Stock, B.C., Jackson, A.L., Ward, E.J., Parnell, A.C., Phillips, D.L., Semmens, B.X.: Analyzing mixing systems using a new generation of bayesian tracer mixing models. PeerJ **6**, 5096 (2018)

[31] Pulley, S., Collins, A.: Tracing catchment fine sediment sources using the new sift (sediment fingerprinting tool) open source software. Science of The Total Environment **635** (2018) https://doi.org/10.1016/j.scitotenv.2018.04.126

[32] Haddadchi, A., Ryder, D.S., Evrard, O., Olley, J.: Sediment fingerprinting in fluvial systems: review of tracers, sediment sources and mixing models. International Journal of Sediment Research **28**(4), 560–578 (2013)

[33] Haddadchi, A., Olley, J., Laceby, J.: Accuracy of mixing models in predicting sediment source contributions. The Science of the total environment **497-498C**, 139–152 (2014) https://doi.org/10.1016/j.scitotenv.2014.07.105

[34] Rowan, J., Goodwill, P., Franks, S.: Uncertainty estimation in fingerprinting suspended sediment sources. In: Tracers in Geomorphology, pp. 279–290 (2000)

[35] Collins, A., Walling, D.: Sources of fine sediment recovered from the channel bed of lowland groundwater fed catchments in the uk. Geomorphology **88**, 120–138 (2007) https://doi.org/10.1016/j.geomorph.2006.10.018

[36] Clarke, R.T.: A bootstrap calculation of confidence regions for proportions of sediment contributed by different source areas in a 'fingerprinting' model. Hydrological Processes **29**(12), 2694–2703 (2015) https://doi.org/10.1002/hyp.10397

[37] Collins, A.L., Zhang, Y., Walling, D.E., Grenfell, S.E., Smith, P., Grischeff, J., Locke, A., Sweetapple, A., Brogden, D.: Quantifying fine-grained sediment sources in the river axe catchment, southwest england: application of a monte carlo numerical modelling framework incorporating local and genetic algorithm optimisation. Hydrological Processes **26**(13), 1962–1983 (2012) https://doi.org/10.1002/hyp.8283 https://onlinelibrary.wiley.com/doi/pdf/10.1002/hyp.8283

[38] Clarke, R.T., Minella, J.P.G.: Evaluating sampling efficiency when estimating sediment source contributions to suspended sediment in rivers by fingerprinting. Hydrological Processes **30**(19), 3408–3419 (2016) https://doi.org/10.1002/hyp.10866

[39] Batista, P.V., Laceby, J.P., Davies, J., Carvalho, T.S., Tassinari, D., Silva, M.L., Curi, N., Quinton, J.N.: A framework for testing large-scale distributed soil erosion and sediment delivery models: Dealing with uncertainty in models and the observational data. Environmental Modelling & Software **137**, 104961 (2021)

[40] Yu, L., Oldfield, F.: A multivariate mixing model for identifying sediment source from magnetic measurements. Quaternary Research **32**(2), 168–181 (1989) https://doi.org/10.1016/0033-5894(89)90073-2

[41] De Maesschalck, R., Jouan-Rimbaud, D., Massart, D.L.: The mahalanobis distance. Chemometrics and intelligent laboratory systems **50**(1), 1–18 (2000)

[42] Daszykowski, M., Kaczmarek, K., Heyden, Y.V., Walczak, B.: Robust statistics in data analysis — a review: Basic concepts. Chemometrics and Intelligent Laboratory Systems **85**, 203–219 (2007)

[43] Minella, J.P., Walling, D.E., Merten, G.H.: Combining sediment source tracing techniques with traditional monitoring to assess the impact of improved land management on catchment sediment yields. Journal of Hydrology **348**(3-4), 546–563 (2008)

[44] Minella, J.P., Merten, G.H., Schlesner, A., Bernardi, F., Barros, C.A., Tiecher,

T., Ramon, R., Evrard, O., Santos, D.R., Reichert, J.M., *et al.*: Combining sediment source tracing techniques with traditional monitoring: The "arvorezinha catchment" experience. Hydrological Processes **36**(9), 14665 (2022)

[45] Davis, J.C., Sampson, R.J.: Statistics and Data Analysis in Geology, p. 646. Wiley, New York (1986)

[46] Harris, C., Millman, K., Walt, S., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S.: Smith 474 nj. Kern R, Picus M, Hoyer S, van Kerkwijk MH, Brett M, Haldane A, del R'ıo JF, Wiebe M, Peterson P, G'erard-475 Marchant P, et al. Array programming with NumPy. Nature **585**(7825), 357–362 (2020) https://doi.org/10.1038/s41586-020-2649-2

[47] Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., *et al.*: Scipy 1.0: fundamental algorithms for scientific computing in python. Nature methods **17**(3), 261–272 (2020) https://doi.org/10.1038/s41592-019-0686-2

[48] Hunter, J.D.: Matplotlib: A 2d graphics environment. Computing in science & engineering **9**(03), 90–95 (2007) https://doi.org/10.1109/MCSE.2007.55

[49] Silva, C., Minella, J., Schlesner, A., Merten, G., Barros, C., Tassi, R., Dambroz, A.: Unpaved road conservation planning at the catchment scale. Environmental monitoring and assessment **193**(9), 595 (2021)

[50] Johnson, R.A., Wichern, D.W., *et al.*: Applied Multivariate Statistical Analysis. Pearson Education Inc, New Jersey (2002)

Figure 1



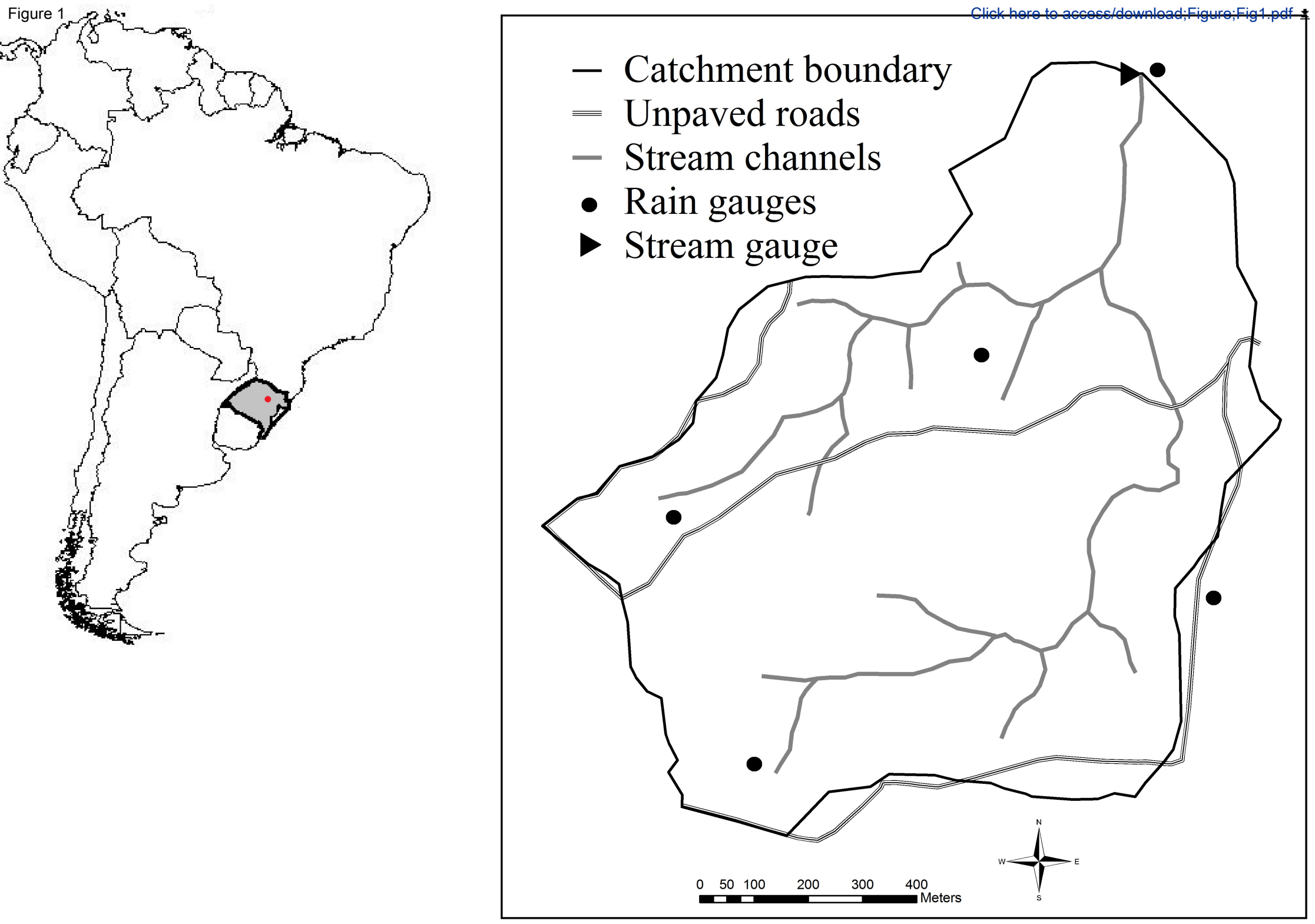


Figure 3a

Click here to access/download;Figure;Fig3a.eps

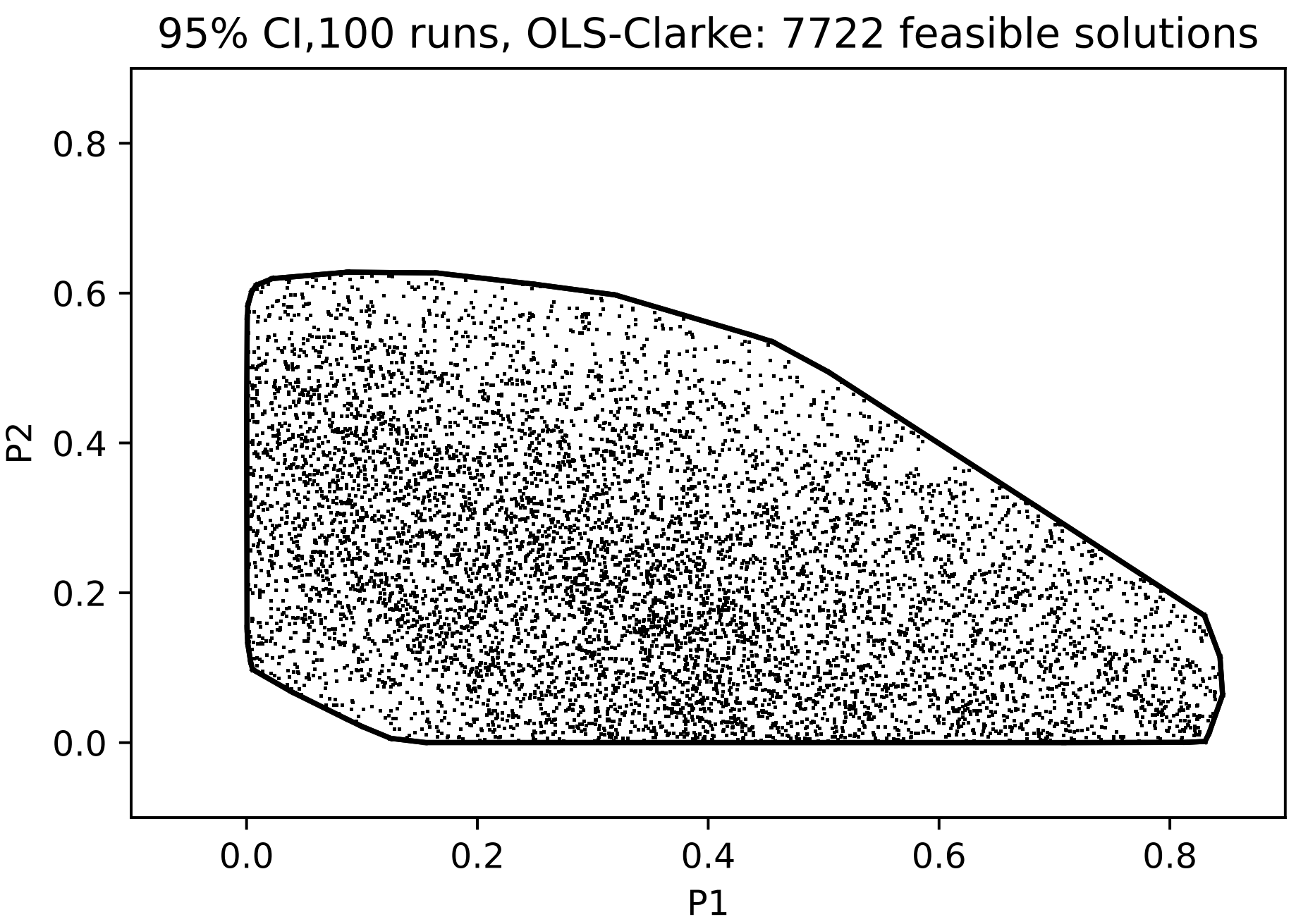


Figure 3b

Click here to access/download;Figure;Fig3b.eps

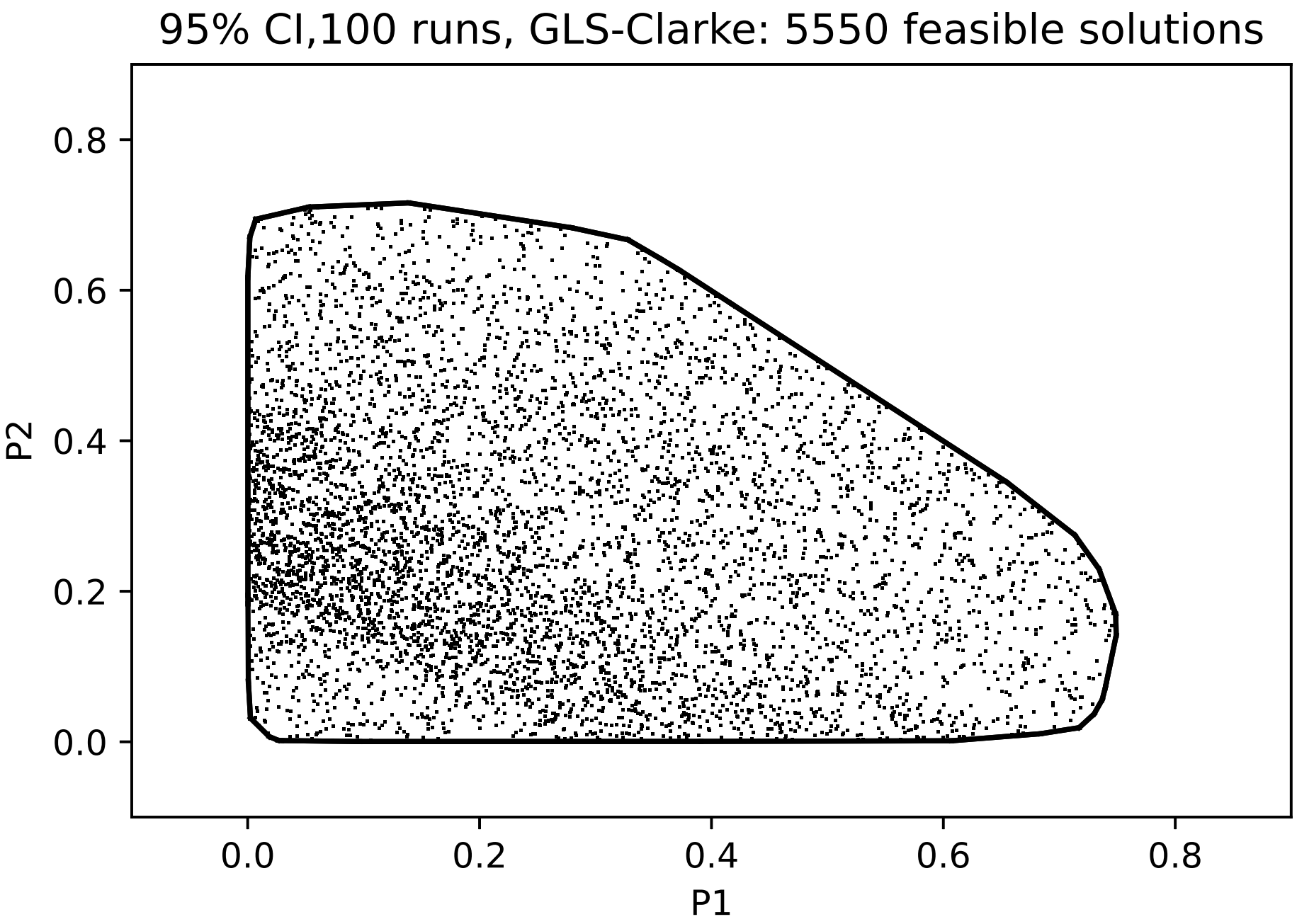


Figure 4a

Click here to access/download;Figure;Fig4a.eps

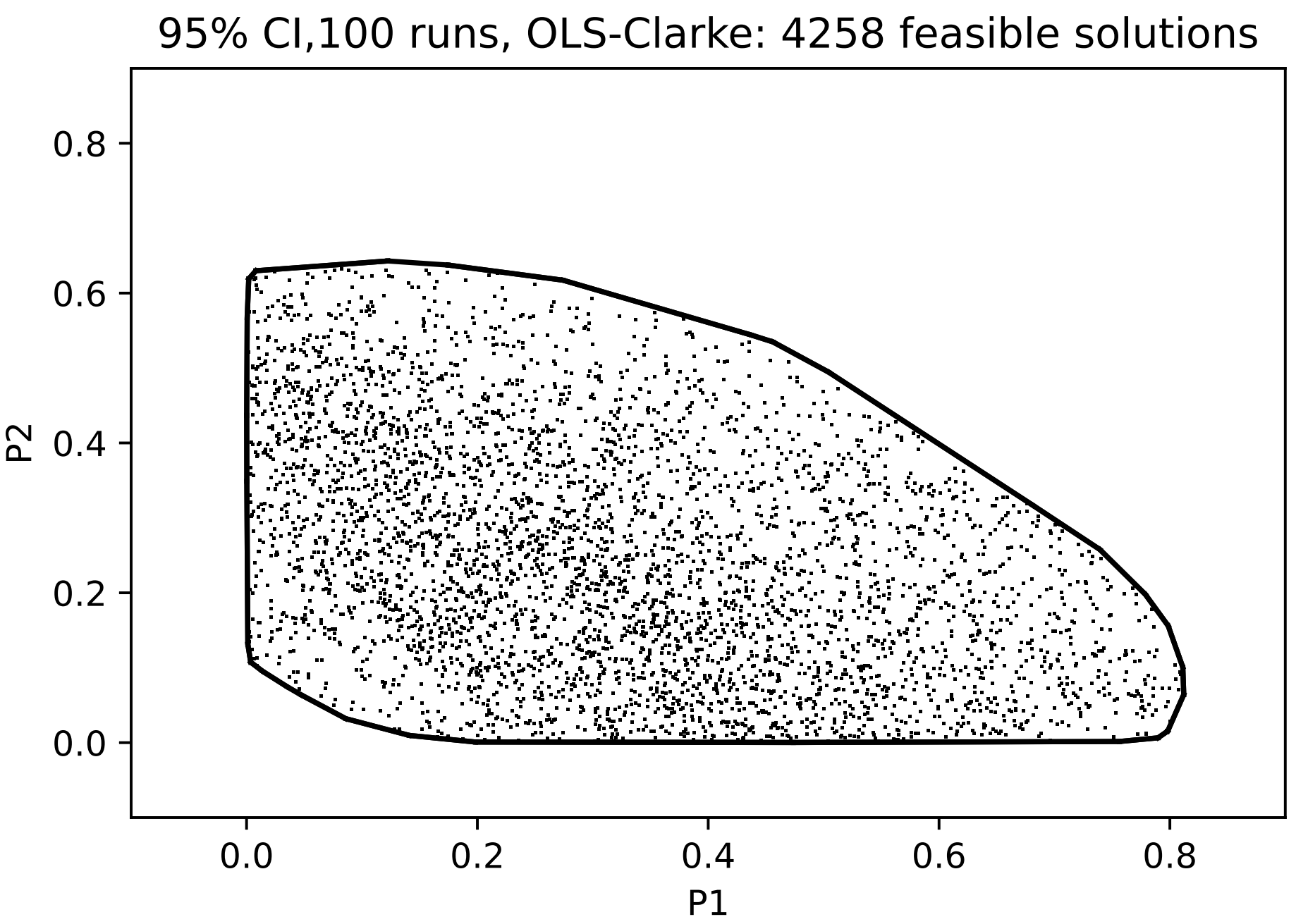


Figure 4b

Click here to access/download;Figure;Fig4b.eps

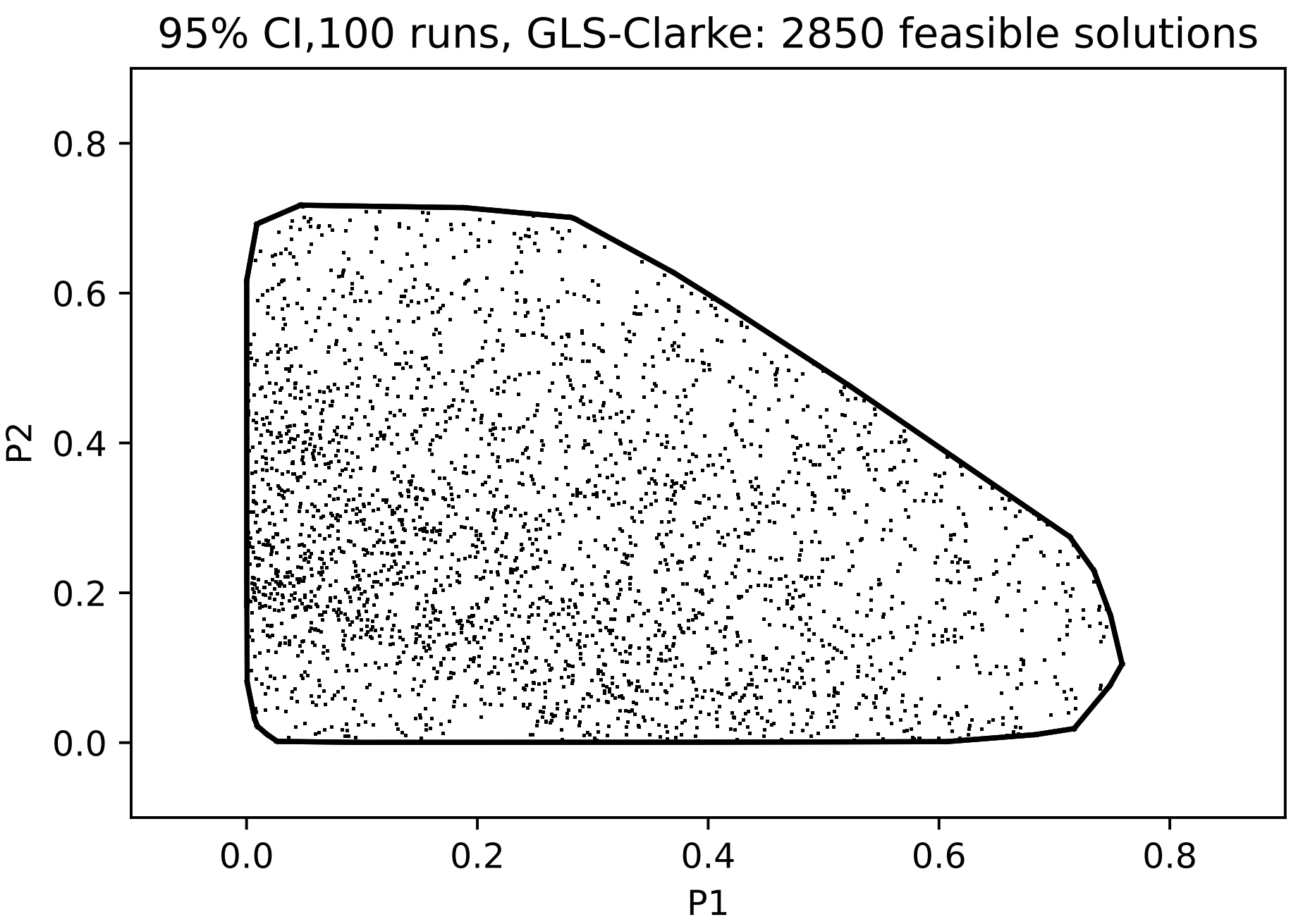


Figure 5a

Click here to access/download;Figure;Fig5a.eps

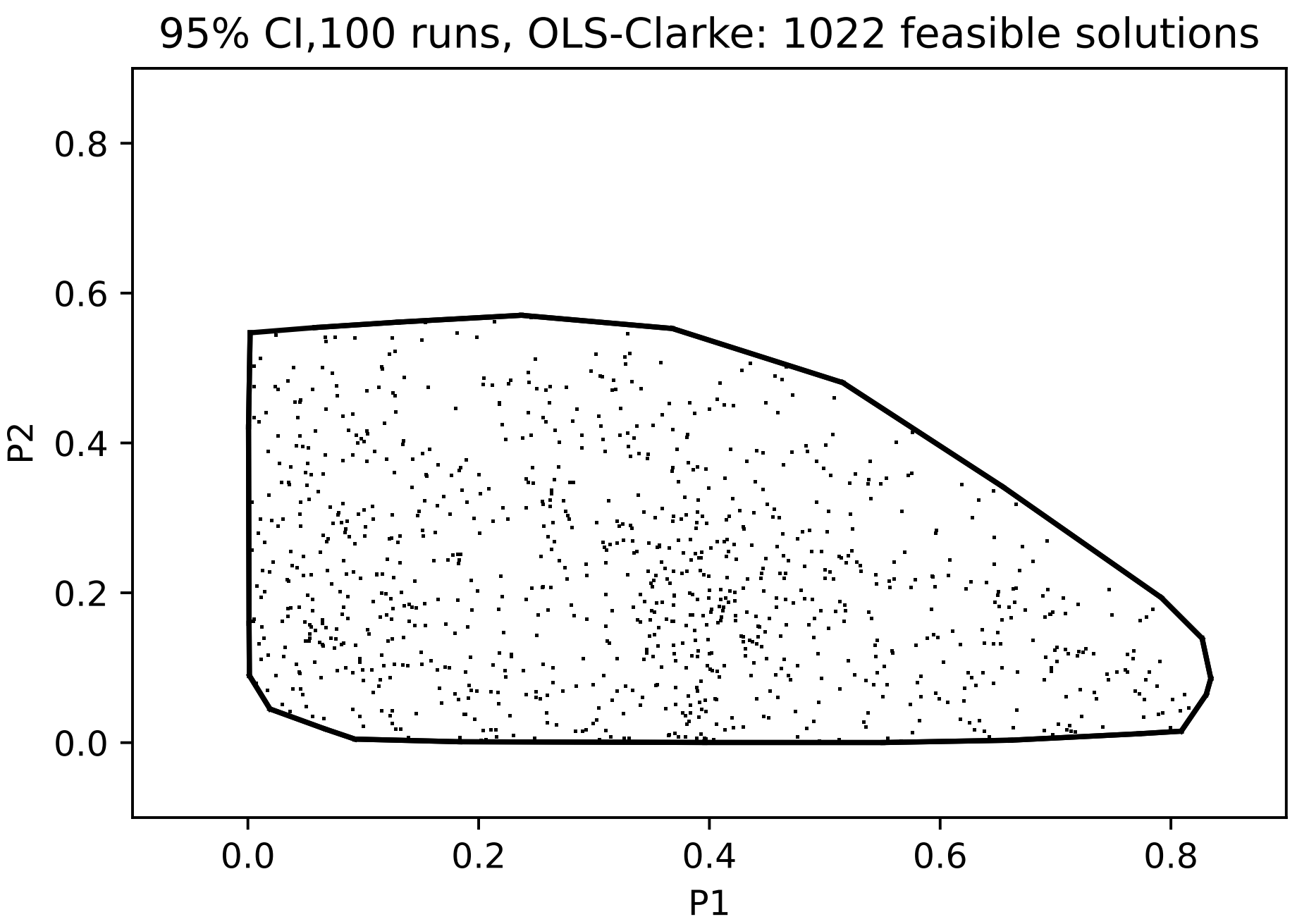


Figure 5b

Click here to access/download;Figure;Fig5b.eps

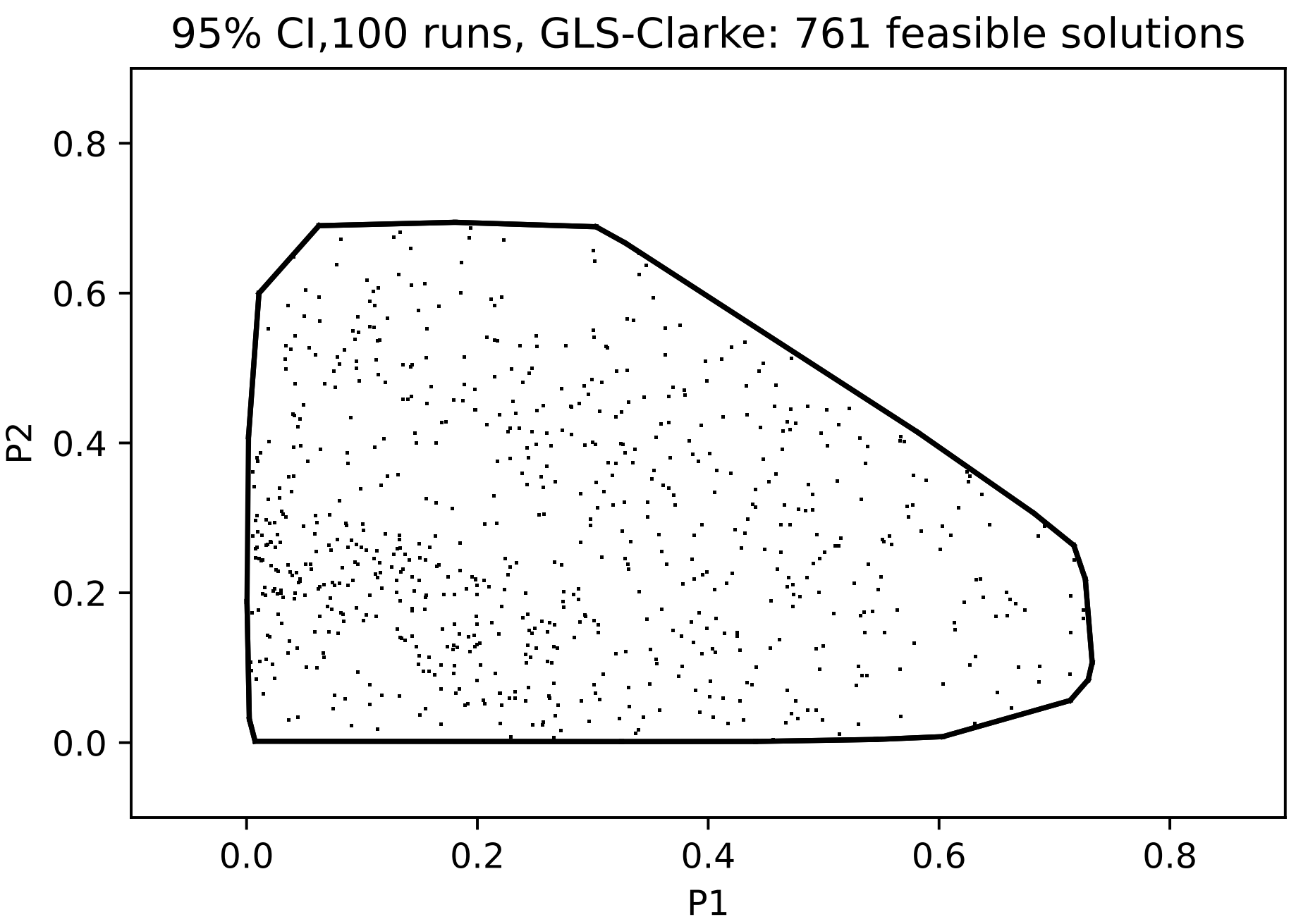


Figure 6

Click here to access/download;Figure;Fig6.eps

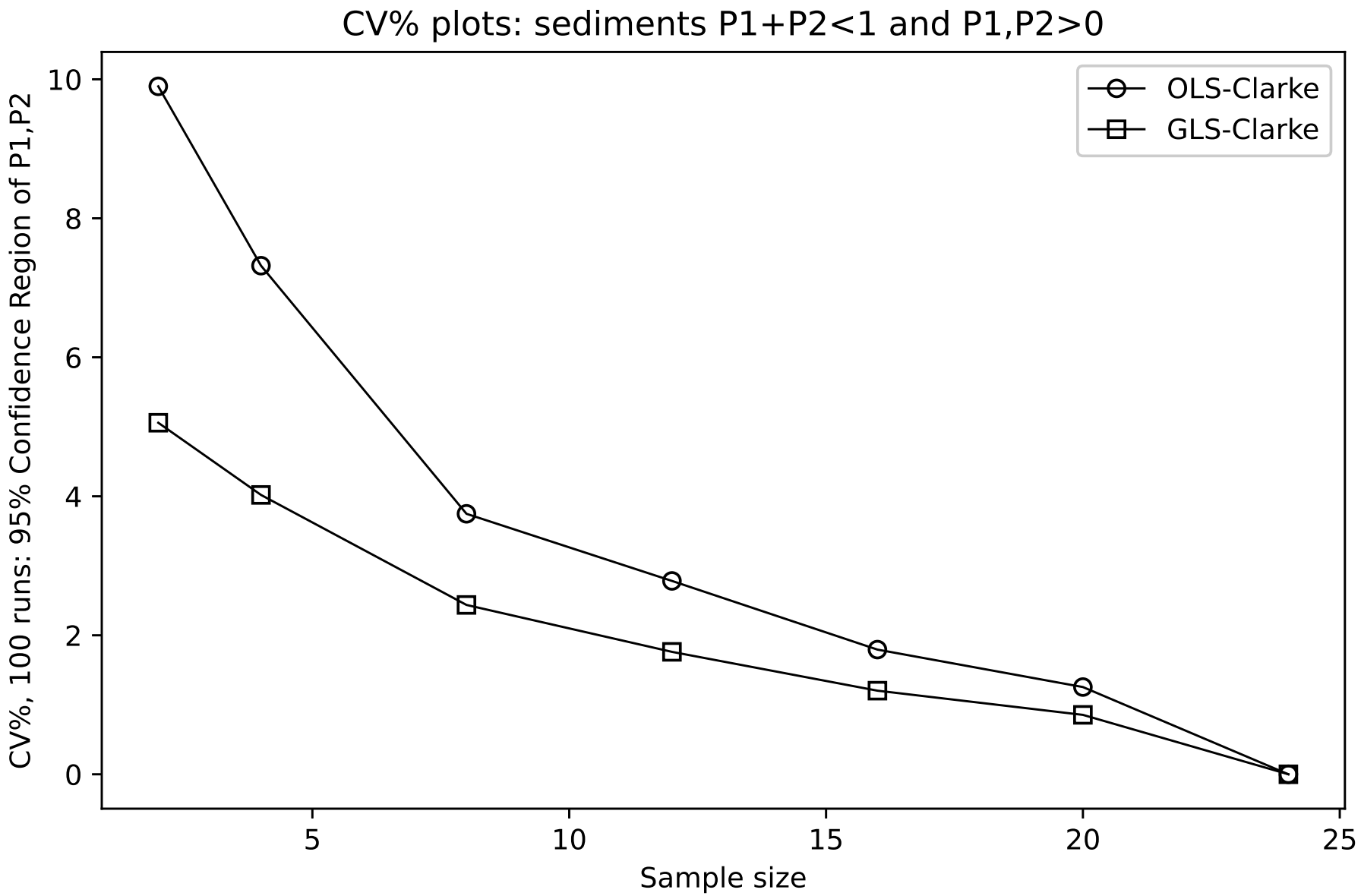


Figure 7a

Click here to access/download;Figure;Fig7a.eps

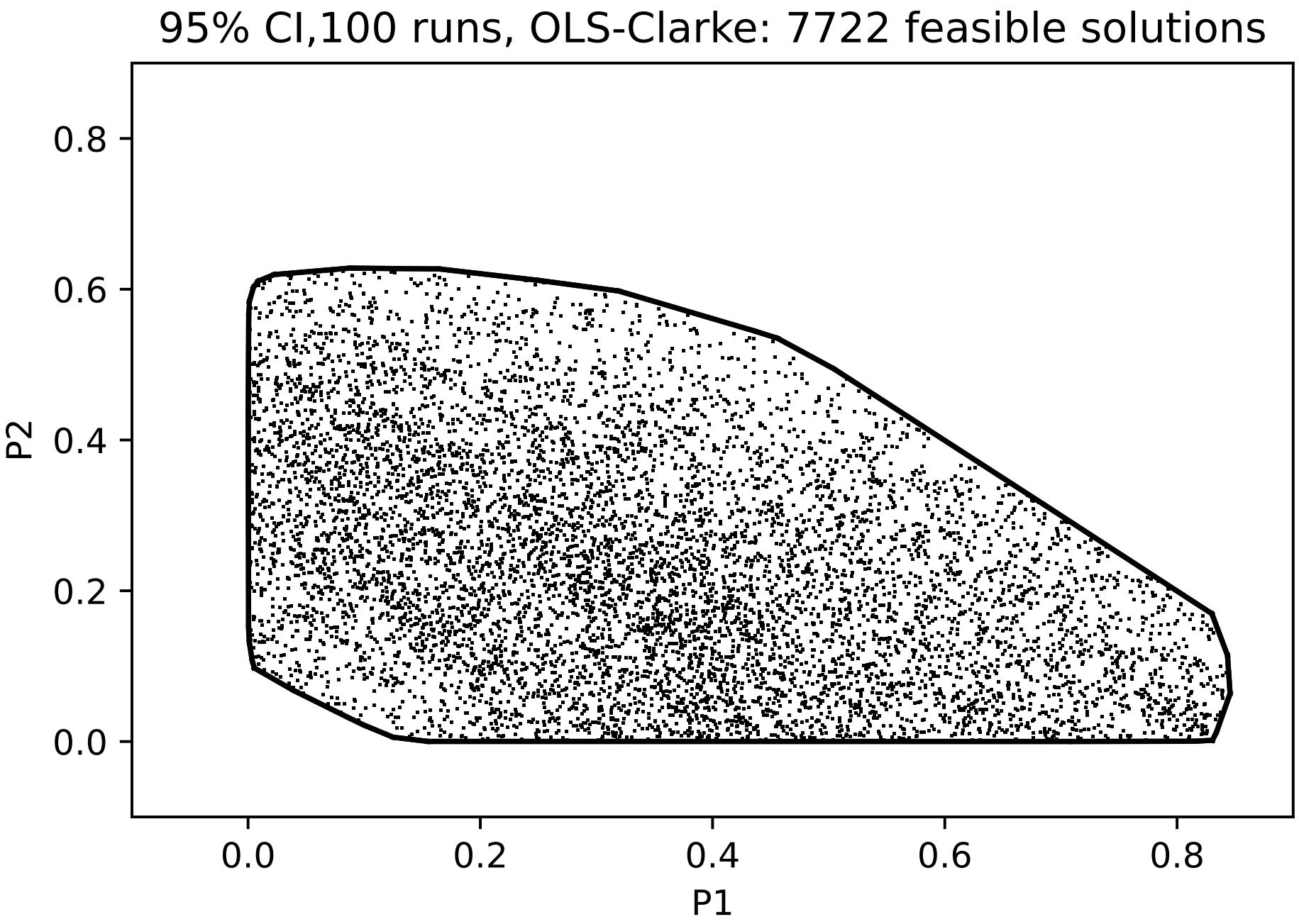


Figure 7b

Click here to access/download;Figure;Fig7b.eps

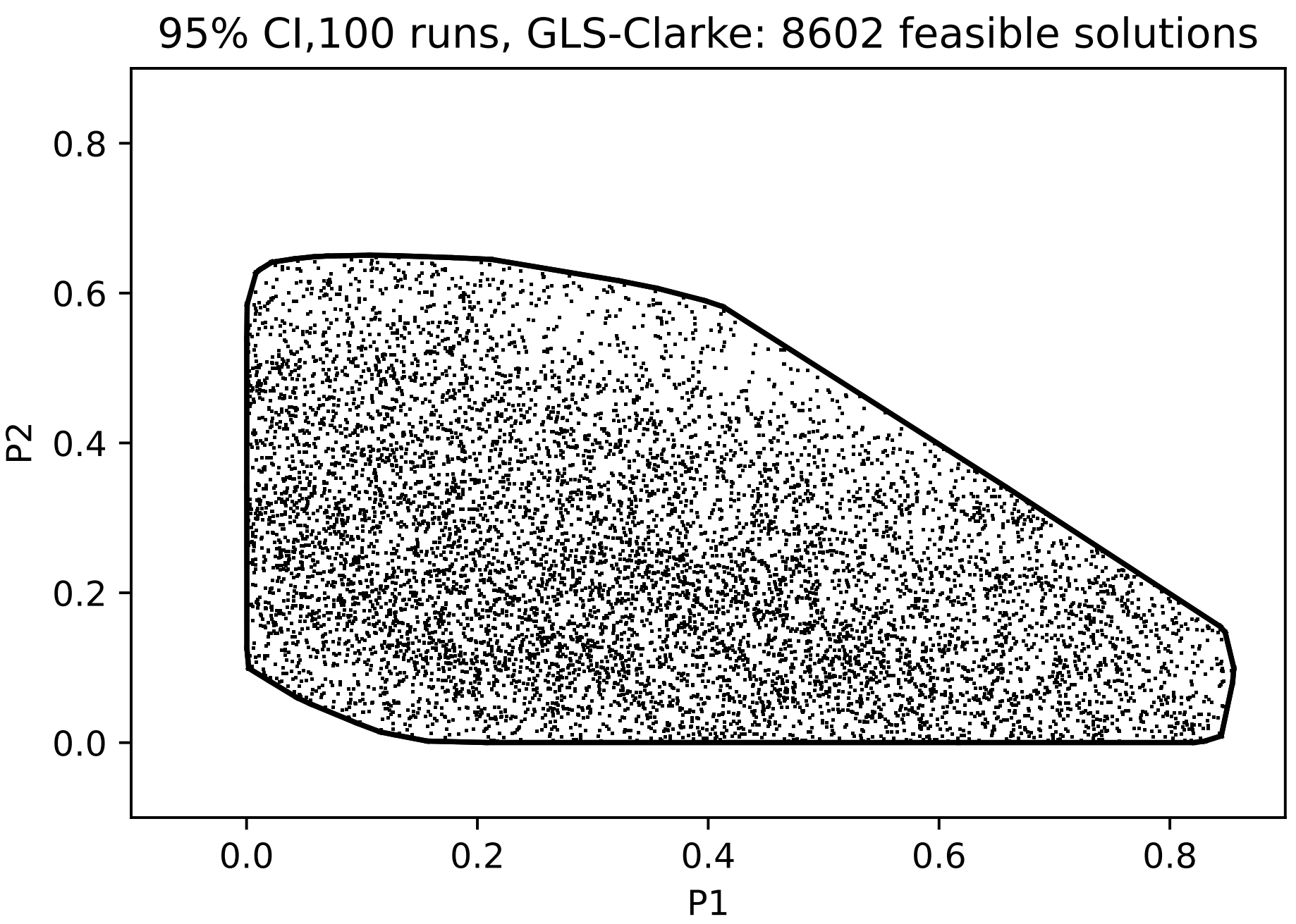


Figure 8a

Click here to access/download;Figure;Fig8a.eps

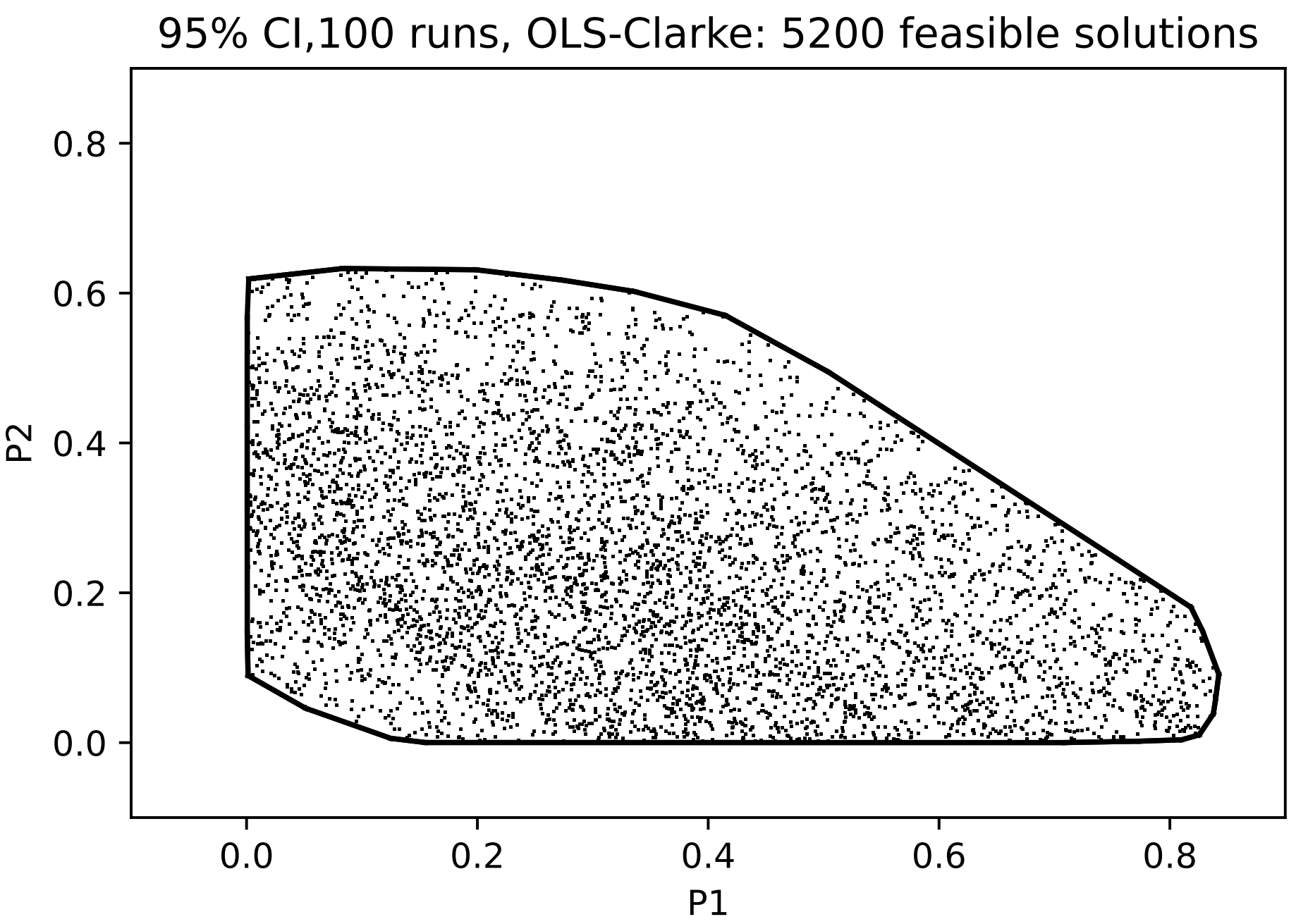


Figure 8b

Click here to access/download;Figure;Fig8b.eps

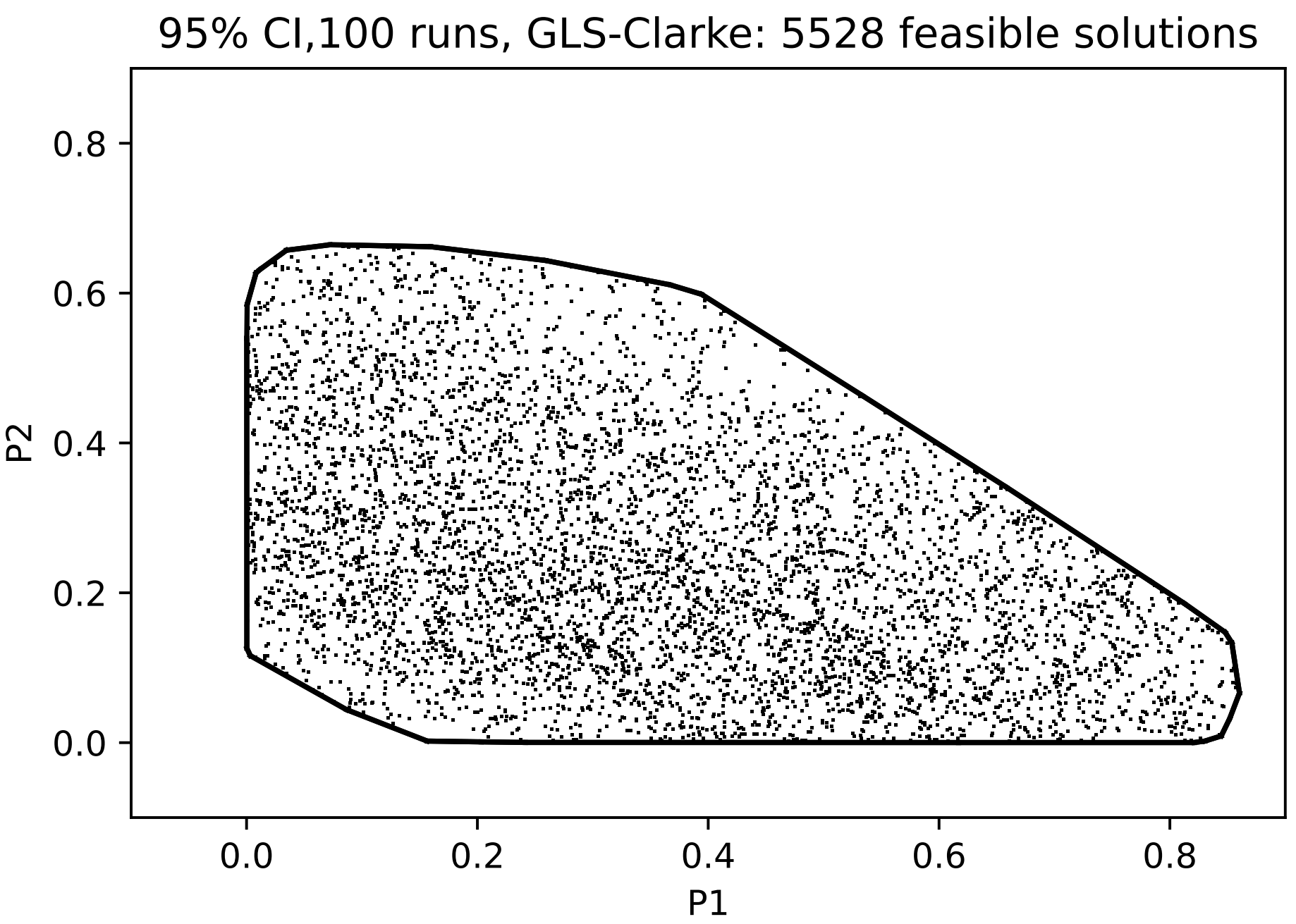


Figure 9a

Click here to access/download;Figure;Fig9a.eps

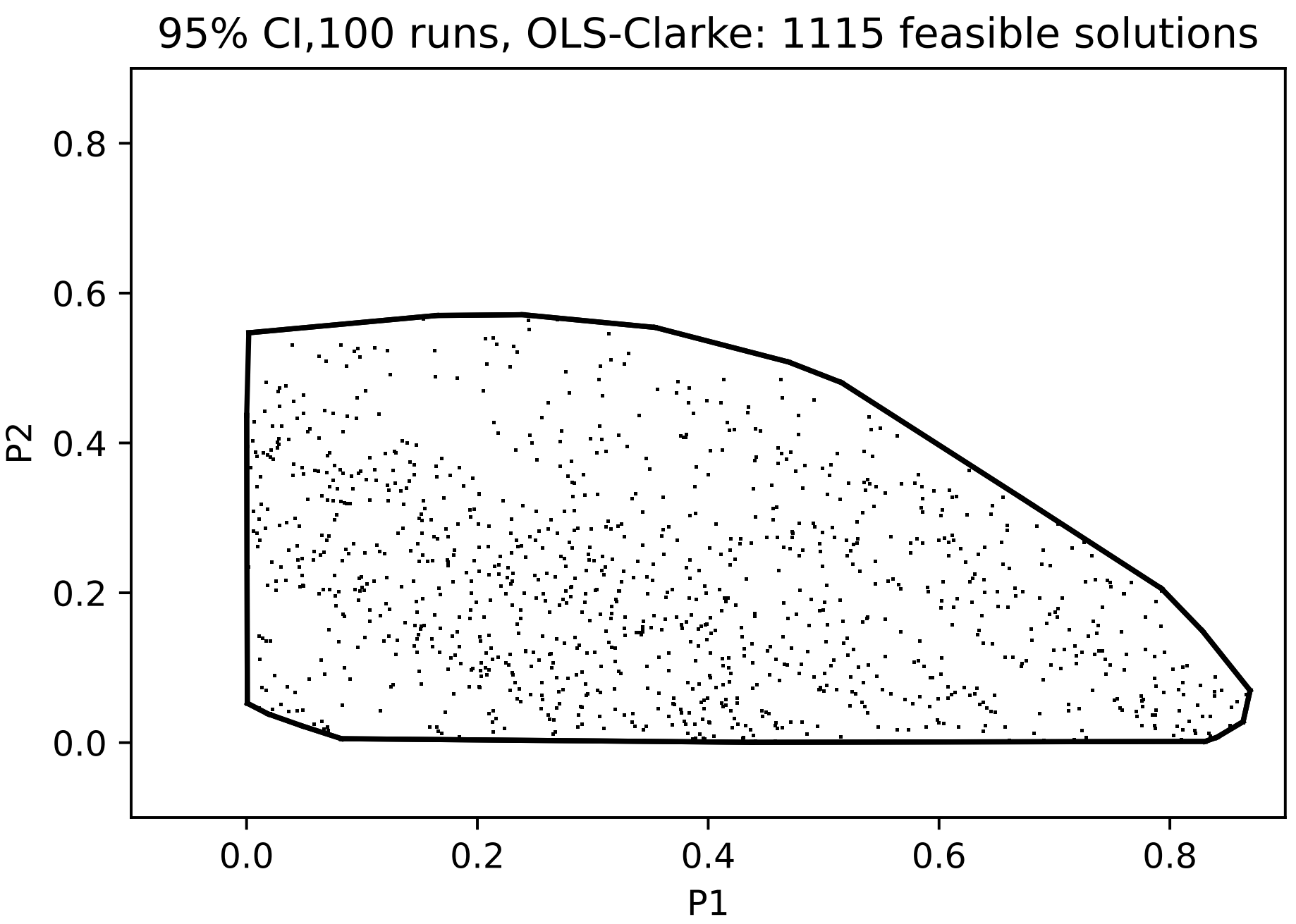


Figure 9b

Click here to access/download;Figure;Fig9b.eps

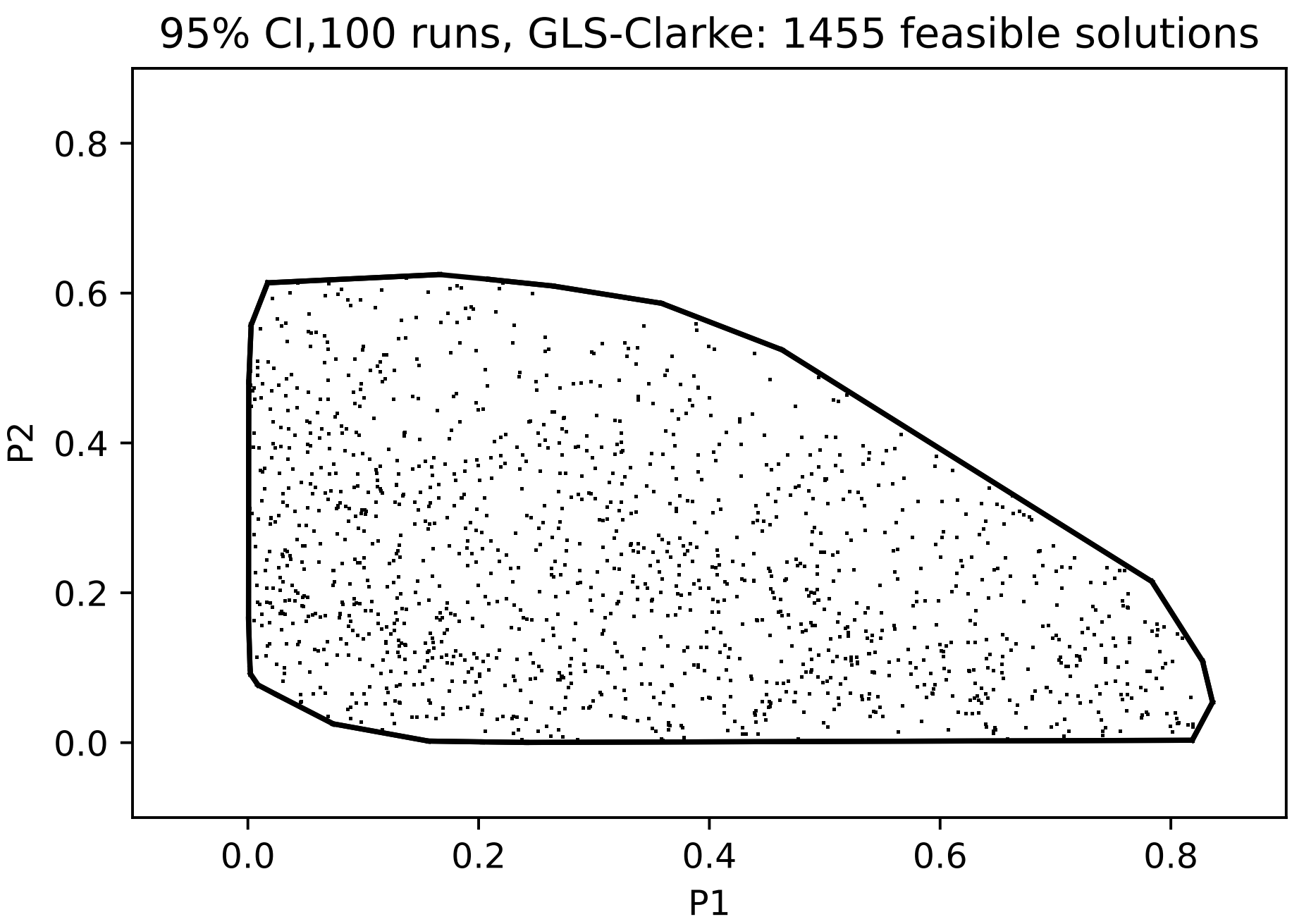


Figure 10

Click here to access/download;Figure;Fig10.eps