entropy

MDPI

Article

On the Information Content of Coarse Data with Respect to the Particle Size Distribution of Complex Granular Media: Rationale Approach and Testing

Carlos García-Gutiérrez [1,*], Miguel Ángel Martín [1] and Yakov Pachepsky [2]

[1] Department of Applied Mathematics, Universidad Politécnica de Madrid, Madrid 28040, Spain; miguelangel.martin@upm.es

[2] USDA-ARS Environmental Microbial and Food Safety Laboratory, Beltsville, MD 20705, USA; Yakov.Pachepsky@ARS.USDA.GOV

* Correspondence: carlos.garciagutierrez@upm.es

Received: 3 June 2019; Accepted: 17 June 2019; Published: 17 June 2019

Abstract: The particle size distribution (PSD) of complex granular media is seen as a mathematical measure supported in the interval of grain sizes. A physical property characterizing granular products used in the Andreasen and Andersen model of 1930 is re-interpreted in Information Entropy terms leading to a differential information equation as a conceptual approach for the PSD. Under this approach, measured data which give a coarse description of the distribution may be seen as initial conditions for the proposed equation. A solution of the equation agrees with a selfsimilar measure directly postulated as a PSD model by Martín and Taguas almost 80 years later, thus both models appear to be linked. A variant of this last model, together with detailed soil PSD data of 70 soils are used to study the information content of limited experimental data formed by triplets and its ability in the PSD reconstruction. Results indicate that the information contained in certain soil triplets is sufficient to rebuild the whole PSD: for each soil sample tested there is always at least a triplet that contains enough information to simulate the whole distribution.

Keywords: information entropy; particle size distribution; selfsimilar measure; simulation

1. Introduction

Granular media resulting from sedimentation and/or fragmentation processes are of great interest in different fields of science, technology and industry. The particle size distribution (PSD) is a main characteristic of these granular media since it has a crucial influence on their physical properties. These media are formed by an enormous amount of particles and, in fact, any particle size within the size interval might potentially be represented in a sample, so that, the PSD may be considered as a continuous distribution. In spite of this, experimental data on this distribution is usually very limited. In the case of soil, a paradigmatic natural granular media, the distribution information is commonly reduced to three classical size fractions, clay, silt and sand [1]. A first natural question arising at this point is if there is a theoretical framework supporting the supposed information value of so limited experimental data. While several mathematical distributions have been used as PSD models [2], the PSD reconstruction from this extremely poor information needs a rationale based in some driving idea different from empirical fitting procedures.

To address this challenge, this work is focused on characterizing the PSD by a specific property which could satisfy a simple equation (differential, difference or dynamical). An outstanding example of this

kind of approach is the pioneering work [3] in which a differential equation is proposed as a semiempirical model for the cumulative mass-size distribution $Q(x)$ of certain granular media with grain size below a given limit x.

$$\frac{\mathrm{d}Q}{\mathrm{d}\,(\log x)} = \alpha Q$$

The differential equation is formulated for granular materials whose grain distribution is arranged in the same statistical manner for both the smaller and for the greater sizes and conformed in such a way that adding a portion of greater grains, the resulting distribution is geometrically similar to the previous one; using the terminology of the authors, they have the same *granulography*.

Interestingly, behind this *old* model swarm features nowadays recognized in many complex dissipative systems. Indeed the formation processes of some granular media have certain aspects that are present in the dynamics of dissipative systems. Fragmentation of particles together with other coupled processes, suggest that the use of energy and its storage takes place in the form of "information" or disorder in the particle sizes. There are two constrains: the available energy for fragmentation is limited and also the energy needed to fragment a particle has a power law dependence of the size of the particle [4]. The maximum entropy principle [5] states that, under certain rules of optimality and randomness, the system thus would reach the maximum level of disorder conditional to the constrains imposed on the process. Entropy maximization methods have already been used to explain the power-scaling nature of size distributions caused by sudden breakage [6]. According to Prigogine [7], the balance of entropy production in dissipative systems should produce a characteristic organization level in a stationary state. In the context of this paper, this corresponds with a characteristic PSD heterogeneity. Notably the term *granulography* used in [3] may be interpreted in a very similar way. These features suggest the use certain elements of Information and Complex Systems Theories in the study of complex granular media, with the goal of establishing a rational basis under which one can evaluate and test the information content of a small number of wide ranges from the distribution.

The paper is organized as follows: in Section 2 a differential information equation is proposed as a conceptual approach for the PSD of complex granular media. Under this approach, experimental data may be seen as initial conditions of the above differential information equation. In Section 3 the use of detailed soil PSD data, together with methods based in the above mathematical approach are used to test the ability of limited experimental data to generate a full reconstruction of the PSD.

2. The Differential Information Equation for the PSD

Instead of the differential equation for the cumulative distribution proposed in [3], we present a rather different type of differential equation framed in a typical quantity used in the description of complex systems: the information entropy (IE).

In mathematical terms, the PSD of granular media may be seen as a mass particle-size distribution μ supported in the interval I of grain sizes.

Limited information on PSD is usually provided as a list of size ranges that cover I. Grains sorted according to their size thus appear distributed in a partition of size classes $P = \{I_1, I_2, \ldots, I_k\}$ defined by those ranges on the list. If the corresponding mass fractions are $p_1 = \mu(I_1), p_2 = \mu(I_2), \ldots, p_k = \mu(I_k)$, respectively, the IE of the partition P is defined by [8]

$$H_\mu(P) = -\sum_{i=1}^{k} p_i \log p_i, \tag{1}$$

provided $p_i \log p_i = 0$ if $p_i = 0$.

The number $H_\mu(P)$ is expressed in information units (bits) and its extreme values are $\log k$, which corresponds to the most even case, when all the intervals have the same cumulative mass; and 0, which corresponds to the most uneven case, when the whole mass is concentrated in a single interval.

The number $H_\mu(P)$ can be interpreted as a measure of heterogeneity. In fact, in [9] it is shown that any measure of heterogeneity having the natural properties for this purpose, must be a multiple of $H_\mu(P)$. Both, the physical hypothesis and the differential equation in [3] have an implicit recognition of scale invariant features. Also the term *granulography* used there agrees with the concept of heterogeneity, which has a precise formulation in mathematical terms, as it was said above. Thus, instead of the differential equation proposed in [3] for the cumulative distribution, we propose a different type of differential equation involving the IE.

If we consider all the partitions $P = \{I_1, I_2, \ldots, I_k\}$ of the size interval I that support the mass particle-size distribution μ, we define

$$H_\mu(r) = \inf\{H_\mu(P) : \mathit{diam}\, P \leq r\} \tag{2}$$

where $r > 0$ and $\mathit{diam}\, P$ is the diameter of P, this is, the length of the greater subinterval of P.

The use of IE allows to formulate a natural property of many multiparticular granular media similar to the master property proposed in [3]: after an arbitrary sieving at a characteristic size scale r, the amount of information received is related to that received at an "inmediately previous" sieving. This relation can be encapsulated in the following initial value problem

$$\begin{cases} \frac{d\, H_{\mu(r)}}{d\, (\log r)} = D, \\ H_\mu(r_0) = H_0 \end{cases} \tag{3}$$

where D is constant and $H_\mu(r_0)$ is the information received at an initial sieving of characteristic size r_0. This is the model we propose for the quantitative description of the PSD.

Although formally this differential equation resembles the proposed in [3], it involves different variables and has a complete different meaning. The physical hypothesis stated by [3], in these new terms, signify that when one travels through the scales using the logarithmic transformation, i.e., changes the size scale, the information content increases on the multiplicative scale. In particular, the equation implies that the information is conserved through the scales.

Under the theoretical point of view, for any partition $P = \{I_1, I_2, \ldots, I_k\}$, the coarse information content of the corresponding empiric data

$$p_1 = \mu(I_1), p_2 = \mu(I_2), \ldots, p_k = \mu(I_k),$$

may be used to provide the initial condition

$$H_0 = -\sum_{i=1}^{k} p_i \log p_i.$$

A first issue is to find out if there is a solution of the Equation (3) corresponding to these initial conditions. Theoretical results from Fractal Geometry [10,11] assure that for each set of empiric data there exists a unique selfsimilar measure, which is a particular solution of (3) using those measured data as initial condition.

Initial data has an static information content that can be calculated with Shannon's entropy. But the information potential of initial data is to suppose that this static information content is mantained, at least

statistically, across the scales, which is exactly what the model (3) implies. Also, there is a scaling behaviour in every natural granular media.

Moreover, it turns out, that in the general case ($p_1 = \mu(I_1)$, $p_2 = \mu(I_2)$, ..., $p_k = \mu(I_k)$) this measure agrees with the proposed as a model for soil mass size particle-size soil distribution in [12]. The latter now appears founded under this approach, and also gives the possibility of simulation. Furthermore, this result links the fractal PSD model proposed in [12] with the model given more than seventy five years earlier in [3].

Nevertheless, the multiplicative cascade associated to each set of experimental data is unique, so a second issue is to study for which initial conditions the corresponding solution μ better reconstructs the real PSD of a certain granular media. This study will take account of which experimental initial conditions (fraction contents) storage greater information content and are most useful to retrieve the actual PSD. There rest of this work is devoted to the study of this problem in the particular case of soil.

3. Materials and Methods

3.1. Data

A total of 70 soil samples from the provinces of Jaen and Segovia in Spain were used. Samples were selected so that they covered the biggest possible part of the USDA textural triangle (Figure 1).

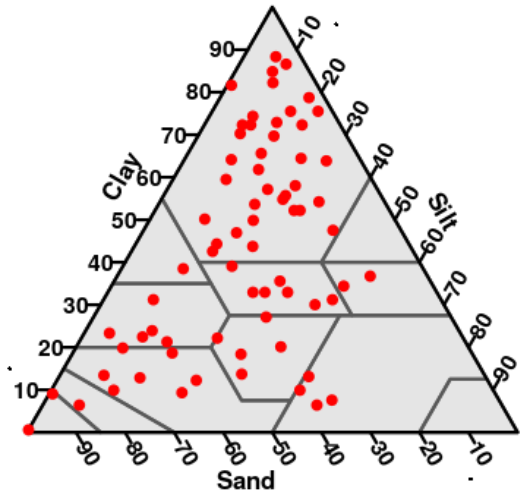

Figure 1. Representation of the 70 soil samples in the USDA textural triangle

They belong to 10 different soil textural classes from the USDA textural classification [1], being the clay class the most represented, with 38 of the samples belonging to it. From a soil classification point of view, these soils belong to to 10 different soil classes, being Calcic Cambisol the most frequent. A complete description of these soils can be found in [13] and references within.

The particle size distribution of these samples was measured using the laser diffraction method [13] with the Longbench Mastersizer S (Malvern Instruments) with a He-Ne laser of 5 mW and a wavelength of 632.8 nm. This apparatus yields a set of data of the form

$$\{I_i = [\varphi_i, \varphi_{i+1}], v_i\}_{i=1}^{N},$$

where N is the total number of size intervals I_i, and v_i is the percentage of total volume of particles whose sizes belong to the size interval I_i. Sizes are given in μm. Let I be the total size interval, i.e., $I = \cup_{i=1}^{N} I_i$.

In this case $I = [0.59\ -\ 3473.45]$ Assuming a constant particle density the probability associated to each interval I_i can be calculated as

$$p_i = \frac{v_i}{\sum_{i=1}^{N} v_i}.$$

The length of the size intervals, I_i, was not constant. The first interval is 0.12 µm and the last one is 574.36 µm. Nevertheless, when using a logarithmic scale, the interval sizes become even and the endpoints of the intervals verify that the quotient $\log \frac{\varphi_{i+1}}{\varphi_i}$ remains equal. Thus, we considered the following size intervals instead:

$$\{J_i = [\phi_i, \phi_{i+1}]\}_{i=1}^{N}, \quad \phi_j = \log_{10} \varphi_j, \qquad j = 1, 2, \ldots, N+1,$$

and J is the new total size interval in the logarithmic scale.

3.2. Simulation and Testing

Soil is a paradigmatic essentially complex granular system whose PSD is usually given in terms of the mass of only three size fractions, clay, silt and sand. These few data are are used as proxy for deriving many soil properties. Therefore it seemed natural to test sets of triplets of intervals along with their masses as initial conditions for the equation, simulate the whole distribution and compare it to a detailed description of the PSD. Indeed, different initial conditions lead to different PSD reconstructions. Testing differrent triplets would allow to find out if any of the them contains enough information to recover the whole PSD.

3.2.1. Triplet Description

The hypothesis was tested using only three mass size intervals in the input partition.

By collapsing the detailed PSD description obtained through the experimental analysis different triplets of mass-size intervals were obtained to use as input data. With 48 available data intervals $\{J_1, \ldots, J_{48}\}$ along their respective masses $\{q_1, \ldots, q_{48}\}$ $(\sum q_i = 1)$, the number of possible combinations of those intervals into a valid triplet $\{T_1, T_2, T_3\}$ was 1081. The mass of each input interval is the sum of all the masses from the data intervals that it comprises.

As an example, let $T_1 = J_1 \cup J_2$, $T_2 = J_3$ and finally $T_3 = \cup_{i=4}^{64} J_i$. Thus, the corresponding masses are $p_1 = q_1 + q_2$, $p_2 = q_3$ and $p_3 = \sum_{i=4}^{64} q_i$.

The geometric description of the triplets is the three intervals it comprises, i.e., $T_1 = [a, b]$, $T_2 = [b, c]$ and $T_3 = [c, d]$. As a and d are the same for all triplets, the transformation

$$\alpha_1 = \frac{b-a}{d-a}, \quad \alpha_2 = \frac{c-a}{d-a},$$

allows for a succint representation of any triplet in the plane. The values of α_1 ranged from 3.504×10^{-5} to 0.697. The values from α_2 ranged from 7.375×10^{-5} to 0.835.

3.2.2. Simulation Algorithm

To simulate the distribution using limited inputs, an iterated function system (IFS) was used [12]. This algorithm can simulate the mass within any interval $A \subset J$, being I the size interval provided by the laser diffraction analysis.

Once the initial interval is divided into the three size fractions that are going to be used as inputs: T_i, where $J = \cup_{i=1}^{3} T_i$, with their respective masses p_i, we shall calculate the linear transformations, ξ_i, $i = 1, 2, 3$, that map J into T_i. Then the simulation algorithm is as follows:

1. take any $x_0 \in I$ as a starting point,

2. choose randomly, with probability p_i, one of the three linear transformations ξ_i, $i = 1, 2, 3$ and calculate the next point in the simulation $\xi_i(x_0) = x_1$,
3. repeat step (2), obtaining a sequence $\{x_k\}$ with $x_k = \xi_i(x_{k-1})$ with probability p_i, chosen randomly, $i = 1, 2, 3$.

This process defines a limit measure. The measure of any interval $A \subset J$, $\mu(A)$ can be calculated as

$$\mu(A) = \lim_{n \to \infty} \frac{m(n)}{n+1},$$

$m(n)$ being the number of points of the orbit $\{x_i\}$ of n points that fall within the interval A.

The simulated distribution was statistically compared to the experimental one using a Kolmogorov-Smirnov (KS) test [14], and thus the distributions are statistically similar.

The convergence of the algorithm is fast. We performed simulations using increasing powers of ten points in the simulation. Only 0.34% of the triplets had a different KS test result when changing from 10^5 to 10^6 points in the simulation.

4. Results and Discussion

There was total of 1081 possible input triplets to simulate the whole PSD. At least 28 triplets (2.6%) passed the KS test for all soils in the database. On average 16.9% (∼182) of the available triplets passed the test for each soil. The number of triplets that passed the test was above 400 (>37%) for soils labeled 6 and 44. Example of simulation results is shown in Figure 2. Simulation with two different triplets is shown. The first one with input intervals $I_1 = [0.59 - 1.46]$, $I_2 = [1.46 - 1406.77]$ and $I_3 = [1406.77 - 3473.45]$ in μm. The values of α_1 and α_2 for this triplet were 0.0003 and 0.4049.

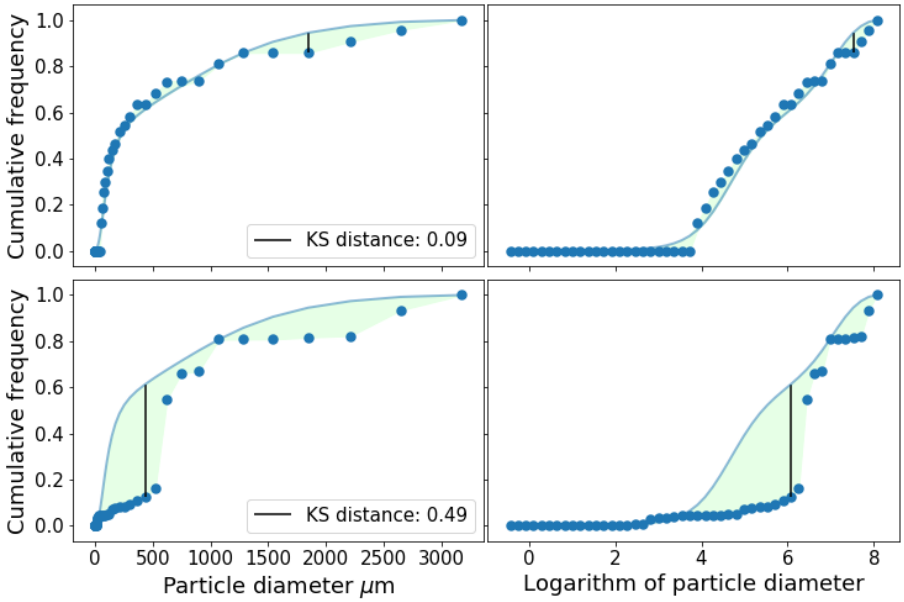

Figure 2. Actual (continuous line) and simulated (dots) PSD for soil 44. Top row shows the simulation using triplet $I_1 = [0.59 - 1.46]$, $I_2 = [1.46 - 1406.77]$ and $I_3 = [1406.77 - 3473.45]$ µm. On the left the x scale is the particle diameter in µm, while on the right, for visualization purposes, it is on the logarithmic scale. For the bottom row, the input triplet used was $I_1 = [0.59 - 37.84]$, $I_2 = [37.84 - 1174.13]$ and $I_3 = [1174.13 - 3473.45]$ µm. In both cases, the maximum allowed distance for the acceptance of the KS test at a 0.05 level was 0.28.

The other triplet was $I_1 = [0.59 - 37.84]$, $I_2 = [37.84 - 1174.13]$ and $I_3 = [1174.13 - 3473.45]$ μm, with $\alpha_1 = 0.0107$ and $\alpha_2 = 0.3379$.

Figure 3 shows the Kolmogorov-Smirnov statistics for all triplets represented in the (α_1, α_2) plane. The horizontal plane at the height of acceptance of the null hypothesis is shown.

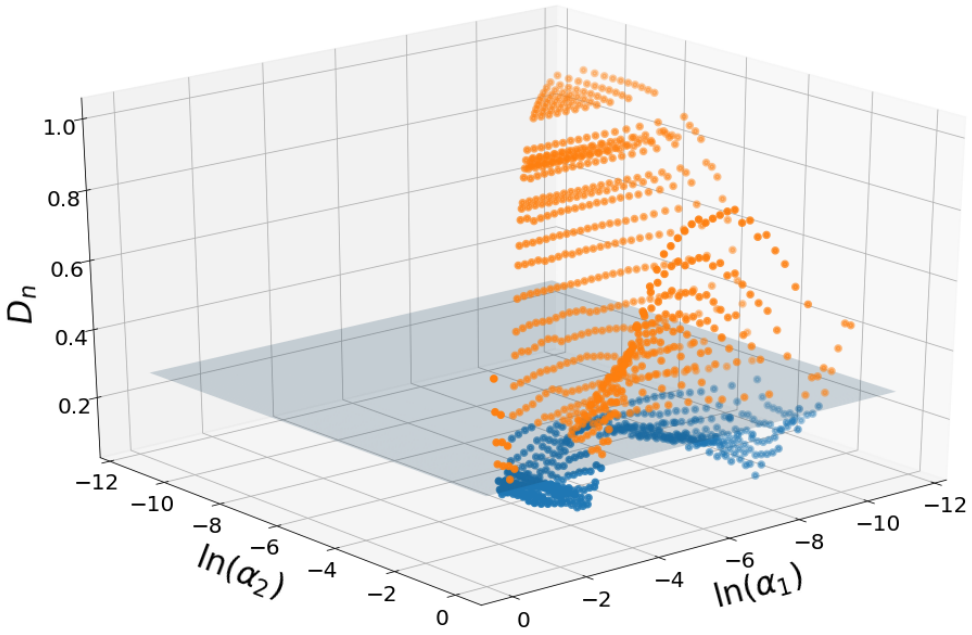

Figure 3. Representation of the KS distance, D_n, for all possible triplets, in the (α_1, α_2) plane, for soil 44. The values of α_i are in the log scale. The horizontal plane, at height, 0.28, is the limit value for D_n for the acceptance region at the 0.05 level. Blue points, below the plane, are the ones that passed the test, while orange ones do not pass it. For this soil, 403 triplets (37.28%) pass the test.

The surface has butterfly-like shape with maximum values found at the extremes of the α_2 values and intermediate α_1 values. Either fine fraction (small α_2) or coarse fraction (large α_2) have to dominate to provide better modeling results.

Total of 536 (49.6%) of all the 1,081 available input triplets passed the KS test for at least one of the soils. The triplet that passed the KS test for most soils was $I_1 = [0.59 - 192.61]$, $I_2 = [192.61 - 979.85]$, $I_3 = [979.85 - 3473.45]$ μm, with $\alpha_1 = 0.055$ and $\alpha_2 = 0.282$. The number of soils for which this triplet passed the KS test was 65 (93%). In terms of the USDA textural classification, this triplet consists of "clay + silt + fine sand", "medium sand + coarse sand", and "very coarse sand + gravel". Total of 22 triplets passed the KS test for 55 soils (79%) or more. The percentage of soils for which each triplet (α_1, α_2) passes the KS test is shown in Figure 4.

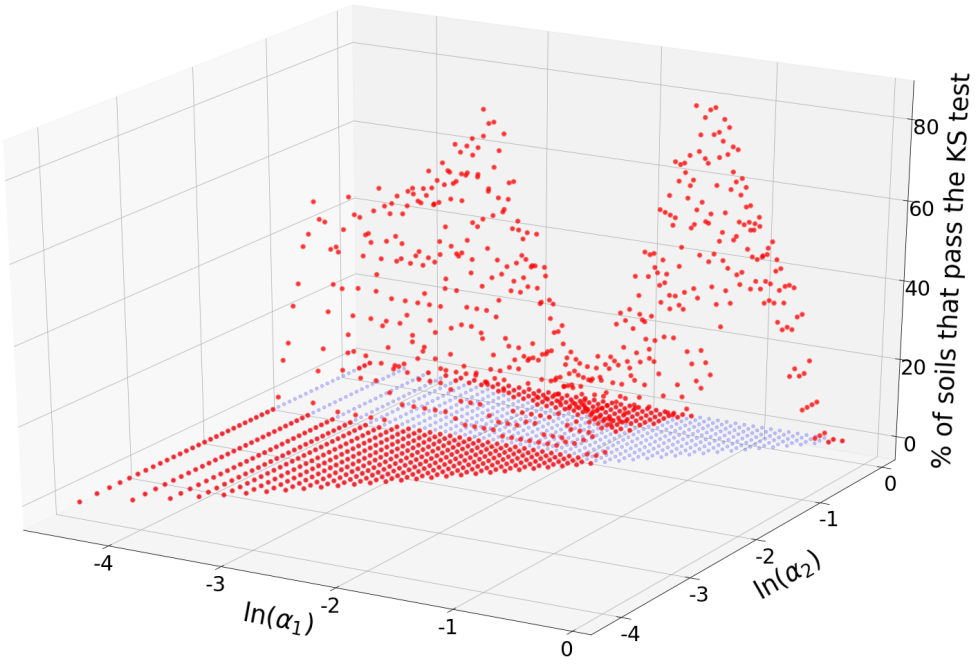

Figure 4. Red dots represent the percentage of samples that pass the KS test for a given triplet (α_1,α_2). For visualization purposes, the projection of the percentages on the plane have been added (blue dots).

Triplets with maximum acceptance rates are concentrated in the neighborhood of the point $(\alpha_1, \alpha_2) = (0.055, 0.282)$, but there are other combinations of alpha values with relatively large acceptances. Figure 5 shows a heatmap of the acceptance percentage for the (α_1, α_2) plane.

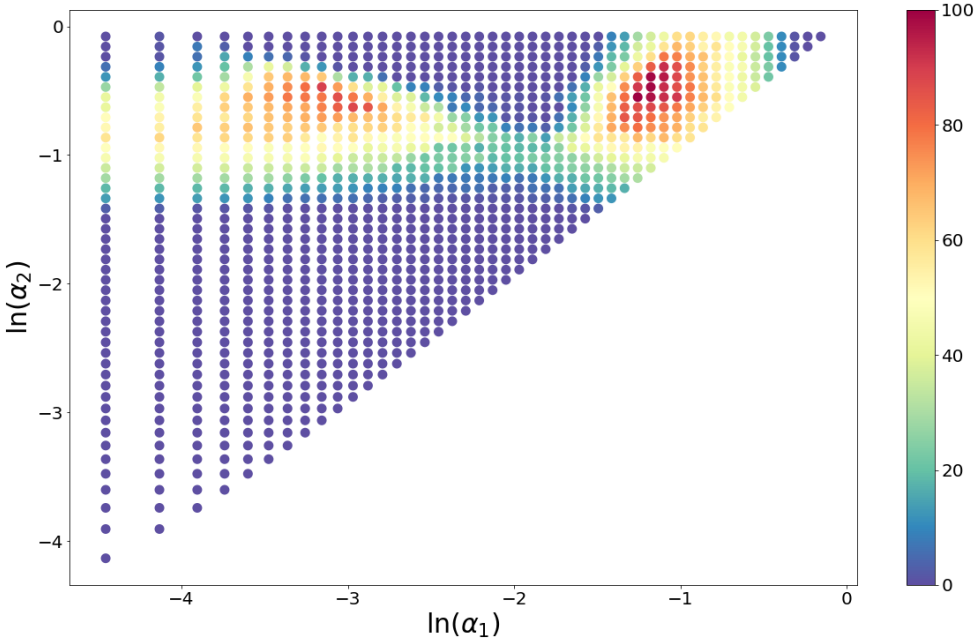

Figure 5. Heatmap for the percentage of samples that pass the KS test for a given triplet (α_1,α_2).

Whether these points attractors have a general significance, are soil database specific or have relations with conditions of soil formation presents an interesting avenue for further research.

5. Conclusions

A new model is proposed for the PSD of granular systems. This model links two works that are separated almost 70 years in time. The link between both models is made by interpreting the driving idea of the first one by means of a differential information equation which leads to the second one. This model provides a theoretical based way to simulate the entire PSD using coarse textural data as initial conditions. This advantage allows to investigate which triplets (as coarse date) stores more information content, this measured by evaluating its potential in reconstructing the real PSD.

Any coarse description of the PSD can be used as an initial condition for the model. In particular, results indicate that the information contained in certain soil triplets is sufficient to reconstruct the whole PSD. For each soil tested there is always at least one triplet that contains enough information to simulate the whole distribution.

Author Contributions: Conceptualization C.G.-G., M.A.M. and Y.P.; methodology, C.G.-G., M.A.M. and Y.P.; software, C.G.-G.; formal analysis, C.G.-G.; investigation, C.G.-G., M.A.M., and Y.P.; validation, C.G.-G.; resources, M.A.M.; data curation, C.G.-G.; writing—original draft preparation, C.G.-G., M.A.M.; writing—review and editing, C.G.-G., M.A.M. and Y.P.; visualization, C.G.-G.; funding acquisition, M.A.M.

Funding: This research work was funded by Spain's Plan Nacional de Investigación Científica, Desarrollo e Innovación Tecnológica (I+D+I), under ref. AGL2015-69697-P.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

PSD	Particle Size Distribution
IE	Information Entropy
KS	Kolmogorov-Smirnov
USDA	United States Department of Agriculture

References

1. Ditzler, C.; Scheffe, K.; Monger, H.C. (Eds.) *Soil Science Division Staff. Soil Survey Manual. Agriculture Handbook 18*; Government Printing Office: Washington, DC, USA, 2017.
2. Bayat, H.; Rastgou, M.; Nemes, A.; Mansourizadeh, M.; Zamani, P. Mathematical models for soil particle-size distribution and their overall and fraction-wise fitting to measurements. *Eur. J. Soil Sci.* **2017**, *68*, 345–364, doi:10.1111/ejss.12423.
3. Andreasen, A.H.M.; Andersen, J. Über die Beziehung zwischen Kornabstufung und Zwischenraum in Produkten aus losen Körnern. *Kolloid-Zeitschrift* **1930**, *50*, 217–228, doi:10.1007/BF01422986.
4. Bertoin, J.; Martínez, S. Fragmentation Energy *Adv. Appl. Probab.* **2005**, *37*, 553–570, doi:10.1239/aap/1118858639.
5. Jaynes, E.T. Information theory and statistical mechanics. *Phys. Rev.* **1957**, *106*, 620-630, doi:10.1103/PhysRev.106.620.
6. Englman, R.; Rivier, N.; Jaeger, Z. Size-distribution in sudden breakage by the use of entropy maximization. *J. Appl. Phys.* **1988**, *63*, 4766, doi:10.1063/1.340114.
7. Prigogine, I. Modération et transformations irreversibles des systèmes ouverts. *Bull. Classe Sci. Acad. R. Belg.* **1945**, *31*, 600–606.
8. Shannon, C.E. A mathematical theory of communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423, doi:10.1002/j.1538-7305.1948.tb01338.x.

9. Khinchin, A.Y. *Mathematical Foundation of Information Theory*; Dover Publications: Mineola, NY, USA, 1957; ISBN 978-048-660-434-3.
10. Deliu, A.; Geronimo, J.S.; Shonkwiler, R.; Hardin, D. Dimensions associated with recurrent self-similar sets. *Math. Proc. Camb. Philos. Soc.* **1991**, *110*, 327–336, doi:10.1017/S0305004100070407.
11. Morán, M.; Rey, J.-M. Singularity of self-similar measures with respect to Hausdorff measures. *Trans. Am. Math. Soc.* **1998**, *350*, 2297–2310 doi:10.1090/S0002-9947-98-02218-1.
12. Martín, M.A.; Taguas, F.J. Fractal modelling, characterization and simulation of particle-size distributions in soil. *Proc. R. Soc. Lond. A* **1998**, *454*, 1457–1468, doi:10.1098/rspa.1998.0216.
13. Montero, E. Aplicacion de Técnicas de Análisis Multifractal a Distribuciones de Tamaño-Volumen de Partículas de Suelo Obenidas Mediante análisis por Difracción de Láser. Ph.D. Dissertation, Universidad Politécnica de Madrid, Madrid, Spain, 2003.
14. DeGroot, M.H. *Probability and Statistics*, 2nd ed.; Addison-Wesley Publ. Co.: Reading, MA, USA, 1986; ISBN 978-013-468-700-1.