# A quick introduction to my research

Daniel Perez

January 12, 2022

## 1 Persistent homology of random functions

My field of study is perhaps better understood by trying to answer the following question :

**Question 1.** How many mountains of height $r$ are there ?

*A priori*, one might think that this is an easy question, since it involves only counting the local maxima of elevation function we encounter and measuring their height. However, given some very rough terrain, the number of these local maxima can be high and any two local peaks can be arbitrarily close to one another, so this definition does not fit the intuitive concept of a "peak". To give a good definition, consider inundating the terrain with water. Given a certain level of the water $r$, we will have a different number of islands or continents, which will vary as we increase $r$. For each $r$, we can confidently say that the number of peaks of height greater than $r$ is exactly equal to the number of continents we count when we have inundated the terrain up to the level $r$. Of course, there will be continents which will get innundated faster than others as the water level rises. The difference in water level between the level at which a new continent appears and the level at which it gets innundated (in technical jargon, its *persistence*) allows us to give a definition for what a *prominent* or *notable* peak is. In this sense, peaks associated with very tall mountains will have very high persistence, while small hills will have low persistence.

A bit more abstractly, we can rephrase the question of the number of peaks as follows. Consider a function $f : X \to \mathbb{R}$ such that $f$ is continuous. We are interested in studying the topology of the sets where the function $f$ takes values greater than some value $r \in \mathbb{R}$. In fact, the topology of these sets can be quite complicated, so we restrict ourselves to studying the so-called homology (which roughly corresponds to a notion of "holes") of these sets. Let us focus on 0th degree homology, which describes the connected components of the level sets and how this homology varies as we vary the level $r$. This corresponds exactly to counting the connected components or "continents" we previously described.

**Remark 1.1.** Looking at higher homology levels amounts to asking whether the continents contain lakes, or holes inside them and if so, how many of them are there ? In higher dimensions, this notion of "holes" can be generalized, leading to a notion of what having an $n$th dimensional hole means.

This is exactly what we study in the field of persistent homology. The word persistent comes from the so-called decomposition theorem, which states that if the function $f$ and the space $X$ are "nice enough", the homology of these sets can be decomposed in terms of $r$, and that the elements in the decomposition "persist" for some time as we vary the level $r$. The invariant obtained through this decomposition is called the barcode of the function $f$. In terms of the mountains we just described, this decomposition retains the birth and death of all of the different continents we observed while innundating the landscape. The term persistence comes from the fact that as we innundate the valley, these continents "persist" or stay alive as we vary the water level.

My research revolves around the persistent homology of random functions. To be exact and asking the reader to indulge the author with a bit of jargon, we can phrase this is precisely as follows.

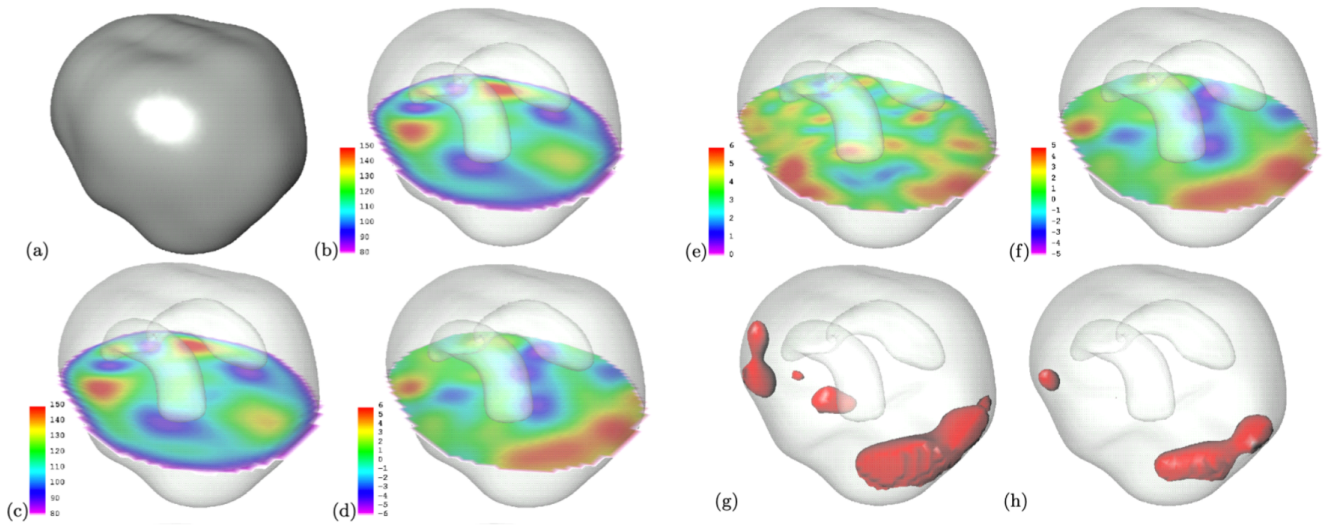


Figure 1: (a) The space $X$, *i.e.* the brain. (b), (c), (d), (e) and (f) are different cross sections of the PET scan. Red zones are more intensely activated. (g) Zones activated above a certain threshhold $r_0$ (h) Zones activated above a higher threshhold $r_1 > r_0$. The figure is taken from [1]

**Question 2.** What can we say about the persistent Betti numbers in expectation (probability, almost surely, *etc.*) of random functions on a compact, connected topological set $X$ ?

This question, while at first hand seemingly abstract, asks very natural questions about the superlevel sets of random functions. This is best illustrated by examples and applications. In dimension one, the question above is equivalent to the following problem.

**Question 3.** How many times does a certain share price fluctuate by at least $p$\$? In particular, how many times does it go up in price from $p_0$ to a price $p_1$ ?

Most of my work is applicable to exactly this question, as most of the results I have thus far obtained concern stochastic processes in dimension one.

Another possible translation of Question 2 in neurobiology has been explored by Adler and Taylor [1]. Here, we are interested in isolating zones of the brain responsible for certain tasks. To do this, we take brain scans using techniques such as positron emission tomography (PET) and see which zones of the brain are activated under some stimuli. The signal we detect takes the form of a real valued function $f$ over the scanned brain, which here will play the role of the topological space $X$, whose values correspond to the level of "activation" at a certain point of the brain (*cf.* figure 1).

We consider a zone of the brain to be activated under the stimuli, if the signal stemming from a PET scan is above a certain threshhold of activation $r$. We can see that it is thus important to quantify the probability that certain regions are activated by chance, so we may determine what the correct value of this threshhold should be. To do this, we must understand the statistical properties of "activated" zones for random functions on $X$. In other words, we would like to know if we could replace the physically meaningful brain activity level measurements by something completely random. If the answer is yes, then, either we set the threshhold $r$ too low, or there is no information in the measurements and some "activated" regions in the scan could be considered as random sets of some kind. Question 2 can thus be rephrased in this context as follows.

**Question 4.** How many areas of the brain do we expect to be "activated" at random in a PET scan, given a threshhold of activation $r$ ? What can we say about their shape ? For instance, do we expect these regions to have holes in them, if so, how many ?

As we can see, questions which can appear at first glance to be quite abstract, quickly becomes concrete when confronted to statistical necessities of certain fields. In light of these few but varied examples, we hope that the reader has convinced him or herself with the necessity to study such questions from a mathematical and abstract point of view.

## 2 Past Research

Random sets and in particular the superlevel sets of random functions have been studied quite extensively in different settings. For example, an abstract version of Question 4 has been treated in Adler and Taylor's celebrated book [1], but also by Azaïs and Wschebor's work [2]. In the settings looked at by these authors, the functions studied are assumed to be smooth, and the authors make conclusions not about particular fixed homology groups we previously aluded to ($n$-degree holes), but rather about invariants combining them, such as the Euler characteristic. This partly answers Question 4, in the sense that as the threshold $r$ grows large, we don't expect the random sets to have holes in them, so the Euler characteristic essentially counts the number of connected components. To avoid taking this limit, results about particular homology groups would be desirable. Other authors to have worked in this setting, asking questions such as the number of points or connected components of a given *level* set, *i.e.* a set where the function $f$ is exactly equal to $r$, include Rice [17] and other authors (*cf.* the references in [1, 2]).

By contrast, in dimension one, this question has been studied in a context where the function $f$ is very irregular, in fact, not even differentiable. In this setting, it is practical to use a construction originally proposed by Le Gall [8], which he later revisited with Duquesne [9], which consists in associating a tree to a function $f$. Formally, this is done by introducing a pseudo-distance defined in terms of $f$ and quotienting out the set of points which are a pseudo-distance 0 away from one another. Informally, it is possible to see this tree appear easily, by considering the graph of some function, and imagining that we use glue along the bottom of the curve, and then push the curve horizontally against the $y$-axis. This is illustrated in figure 2. The resulting object is a tree. Possible extensions to higher dimensions of this construction

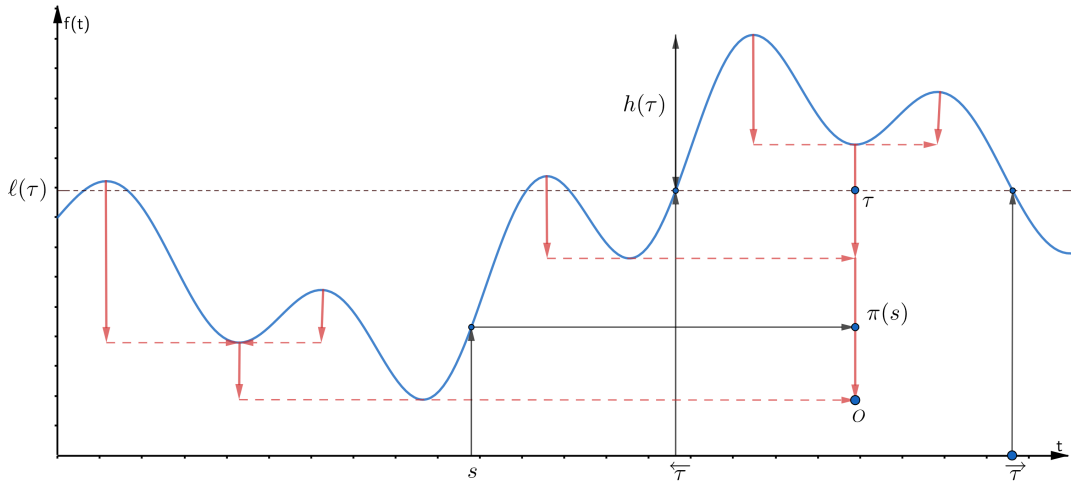


Figure 2: A function $f$ in blue and its associated tree in red. Points connected by dots are identified.

had previously been suggested by Curien *et al.* [6], in which case the "gluing" picture is similar to that of the one-dimensional case. Crucially, these trees have the property that they identify points along the curve belonging to the same level-set lying in the same connected component of the superlevel set.

Much work has been done in the probability theory community regarding these trees, but some particularly useful and key references in the context of my work is the work of Picard [16] and that of Neveu and Pitman [10]. Much of the research done in this direction concerns asymptotics of a certain quantity, which, in the language of Question 3, is the number of fluctuations of size at least $\varepsilon$ in the share

price there are. This can be transposed in terms of the tree as follows. First consider a version of the tree which we have trimmed by $\varepsilon$, that is, we have cut off all pieces which are less than $\varepsilon$ away from a leaf. Let us denote this trimmed tree $T_f^\varepsilon$. Denoting $N^\varepsilon$ the number of the truncated tree $T_f^\varepsilon$, the question of the fluctuations of the share price $f$ can be reinterpreted as finding $N^\varepsilon$. Authors like Picard [16], Neveu and Pitman [10] made significant contributions to answering this question. The former found asymptotics of $N^\varepsilon$ for small $\varepsilon$ and the latter reinterpreted the problem in terms of stopping times, which have been widely studied in the theory of stochastic processes.

My contribution to this probabilistic approach, based on trees, was to show that a tweaked version of this pseudo-distance generates a tree identifying connected components for any continuous function on a connected, locally path-connected and compact topological space $X$ [13]. I also extended Picard's results by using these asymptotics for various purposes, such as estimating $N^\varepsilon$ for sequences of processes whose limits are the processes studied by Picard.

This problem has also been looked at under the scope of persistent homology and persistence theory by a plethora of authors. Some key references are Oudot's [11] and Chazal *et al.*'s [4] textbooks on the topic and the references theirein. Some authors like Chazal and Divol [5] were able to use some classical results to deduce the expectation of the number of points in any rectangle of the so-called *persistence diagram* of $f$ for Brownian motion. In the language of Question 3, this information is equivalent to knowing, in expectation, how many times the share goes up in price from some fixed price $p_0$ to some higher fixed $p_1$. Similarly, Baryshnikov also performed such computations for Brownian motion with a drift [3]. Interestingly, trees akin to those previously defined by the probabilists were rediscovered in the context of persistent homology. Here, the main authors to be cited include (but are not limited to) Wang *et al.* [20], Curry [7], Baryshnikov [3] and Schweinhart [18].

Given the fact that both of these *a priori* different fields introduce trees, one is tempted to ask whether these constructions are equivalent. And indeed, it turns out that so-called *merge trees* correspond exactly to the notion of trees proposed by the probabilists, as I showed in [13], except that the probabilist's notion extends results easily to a wider class of functions and topological spaces. In particular, this comparison and unification of both approaches allowed me to extend Curry's work [7] with less assumptions of regularity, and, in the stochastic setting, to extend and generalize the results of Baryshnikov [3], Chazal and Divol [5] and Picard [16] to much wider classes of processes. We will elaborate on this later on.

## 2.1 My Contribution

With this wider perspective, we can summarize some of my main contributions along the following points.

- The comparison between trees introduced by probabilists and barcodes and an algorithm linking the two. This proves the equivalence of merge trees and the probabilistic construction in the usual setting where merge trees are considered, but extends the framework to any continuous function on any compact, connected and locally path-connected topological space. In less technical jargon, this expands the generality under which these trees had previously been considered, extending it to potentially "uglier" spaces and "rougher" or more irregular functions. In the language of Question 1, this result shows that this theory can be used to describe all sorts of terrain : not only nice and smooth hilly countrysides as was previously known, but also rough, irregular and mountainous terrain. I also showed that it is possible to give an inequality linking the Hölder regularity of the function $f$ (here, this quantifies how rough the terrain is, quantitatively) and the upper-box dimension of the tree associated to $f$. In fact, I also showed this inequality is optimal, by constructing a function saturating the inequality.

- The application of the theory developed for trees to one dimensional random processes. Following and extending works of authors like Picard, Baryshnikov, Divol and Chazal. In the language of Question 3, I studied the number of fluctuations of the share price $N^\varepsilon$ for large values of $\varepsilon$ in expectation, and its higher moments, which was previously not considered in the literature for a

wide range of processes (Markov processes). I found explicit results for Brownian motion and the Brownian bridge, quantifying these large variations precisely [14].

- By approximation, I showed that the results obtained for the canonical processes above can be adapted and yield consequences for any almost surely convergent sequence in $L^\infty$-norm to these processes, in which case we also estimate $N^\varepsilon$ and give a bound on the error of the approximation. Put simply, even if the share price is not Brownian, but is modelled as a discrete random walk with steps of finite variance, this theorem shows that we still have estimates on the number of large fluctuations of the share price, with known error. This extends considerably the number of models we can consider for share prices where we still know something about the number of large fluctuations in expectation [14].

- The introduction of a notion of $\zeta$-functions for stochastic processes (not to be confused with the classical Riemann $\zeta$-function), defined in terms of the persistent homology and showing that these $\zeta$-functions share some points with the classical Riemann $\zeta$-function from number theory. Explicitly, we may express these $\zeta$-functions for stochastic processes as

$$\zeta(p) := \int_0^\infty \mathbb{E}[N^\varepsilon]\, \varepsilon^{p-1}\, d\varepsilon\,.$$

  Furthermore, I showed that these functions admit a meromorphic extension to the whole complex plane, with a unique pole. From these functions, we get dual information on the expectation number of bars of length $\geq \varepsilon$ (in some cases, the information we get is total, *i.e.* we get a series that converges to the expectation) [15].

- The introduction of these $\zeta$-functions allows us to show precise results for $N^\varepsilon$ for $\alpha$-stable processes[1] in both the small and large $\varepsilon$ regime [15]. This is significant, because, as originally pointed out by Benoît B. Mandelbrot, most markets should be modelled on $\alpha$-stable processes [19] and once again extend the variety of models used for share prices.

- A new statistical test for the parameter $\alpha$ for nowhere monotone $\alpha$-stable Lévy processes.

- The introduction of local $\zeta$-functions, from which one can derive information about the local time of the process. In particular, I have deduced a series computing all integer moments[2] of the local time of the Ornstein Uhlenbeck process in expectation [12].

- Explicit calculations of the $\zeta$-functions of several processes, as well as asymptotic series of the expectation of the number of bars of length $\geq \varepsilon$ (and higher order moments for some examples).

# References


[1] R. J. Adler and J. E. Taylor. *Random Fields and Geometry.* Springer New York, 2007.

[2] J.-M. Azaïs and M. Wschebor. *Level Sets and Extrema of Random Processes and Fields.* John Wiley & Sons, Inc., Jul 2008.

[3] Y. Baryshnikov. Time series, persistent homology and chirality. *arXiv:1909.09846*, 2019.

[4] F. Chazal, V. de Silva, M. Glisse, and S. Oudot. *The Structure and Stability of Persistence Modules.* Springer International Publishing, 2016.


[1] A distribution is said to be $\alpha$-stable if any linear combination of two independent random variables with this distribution has the same distribution. A process is said to be $\alpha$-stable if at every point it is distributed with al $\alpha$-stable distribution, among other technical conditions.

[2] The non-integer case is in principle possible, but technically difficult and I have not included it in the paper.

[5] F. Chazal and V. Divol. The density of expected persistence diagrams and its kernel based estimation. In B. Speckmann and C. D. Tóth, editors, *34th International Symposium on Computational Geometry (SoCG 2018)*, volume 99 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 26:1–26:15, Dagstuhl, Germany, 2018. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

[6] N. Curien, J.-F. Le Gall, and G. Miermont. The Brownian cactus I. scaling limits of discrete cactuses. *Ann. Inst. H. Poincaré Probab. Statist.*, 49(2):340–373, 05 2013.

[7] J. Curry. The fiber of the persistence map for functions on the interval. *Journal of Applied and Computational Topology*, 2(3-4):301–321, dec 2018.

[8] T. Duquesne and J.-F. Le Gall. *Random trees, Lévy processes and spatial branching processes.* Number 281 in Astérisque. Société mathématique de France, 2002.

[9] T. Duquesne and J.-F. Le Gall. Probabilistic and fractal aspects of Lévy trees. *Probability Theory and Related Fields*, 131(4):553–603, Nov 2004.

[10] J. Neveu and J. Pitman. Renewal property of the extrema and tree property of the excursion of a one-dimensional Brownian motion. *Séminaire de probabilités de Strasbourg*, 23:239–247, 1989.

[11] S. Y. Oudot. *Persistence Theory - From Quiver Representations to Data Analysis*, volume 209 of *Mathematical surveys and monographs.* American Mathematical Society, 2015.

[12] D. Perez. Local $\zeta$-functions in stochastic persistent homology.

[13] D. Perez. On $C^0$-persistent homology and trees. https://arxiv.org/abs/2012.02634, Dec. 2020.

[14] D. Perez. On the persistent homology of almost surely $C^0$ stochastic processes. https://arxiv.org/abs/2012.09459, Dec. 2020.

[15] D. Perez. $\zeta$-functions and the topology of superlevel sets of stochastic processes. *arXiv e-prints*, page arXiv:2110.10982, Oct. 2021.

[16] J. Picard. A tree approach to $p$-variation and to integration. *The Annals of Probability*, 36(6):2235–2279, Nov 2008.

[17] S. O. Rice. Mathematical analysis of random noise. *The Bell System Technical Journal*, 23(3):282–332, 1944.

[18] B. Schweinhart. Persistent homology and the upper box dimension. *Discrete Computational Geometry*, Nov 2019.

[19] J. Voit. *The Statistical Mechanics of Financial Markets.* Springer-Verlag, 2005.

[20] S. Wang, Y. Wang, and R. Wenger. The JS-graphs of join and split trees. In *Proceedings of the thirtieth annual symposium on Computational geometry.* ACM, jun 2014.