Functional Kernel Density Estimation: Point and Fourier Approaches to Time Series Anomaly Detection

Michael R. Lindstrom [1,*] , **Hyuntae Jung** [2] **and Denis Larocque** [3]

[1]   Department of Mathematics, University of California, Los Angeles, CA 90024, USA
[2]   Global Aviation Data Management, International Air Transport Association (IATA),
    Montréal, QC H2Y 1C6, Canada; jungh@iata.org
[3]   Department of Decision Sciences, HEC Montréal, Montréal, QC H2Y 1C6, Canada; denis.larocque@hec.ca
[*]   Correspondence: mikel@math.ucla.edu

Abstract: We present an unsupervised method to detect anomalous time series among a collection of time series. To do so, we extend traditional Kernel Density Estimation for estimating probability distributions in Euclidean space to Hilbert spaces. The estimated probability densities we derive can be obtained formally through treating each series as a point in a Hilbert space, placing a kernel at those points, and summing the kernels (a "point approach"), or through using Kernel Density Estimation to approximate the distributions of Fourier mode coefficients to infer a probability density (a "Fourier approach"). We refer to these approaches as Functional Kernel Density Estimation for Anomaly Detection as they both yield functionals that can score a time series for how anomalous it is. Both methods naturally handle missing data and apply to a variety of settings, performing well when compared with an outlyingness score derived from a boxplot method for functional data, with a Principal Component Analysis approach for functional data, and with the Functional Isolation Forest method. We illustrate the use of the proposed methods with aviation safety report data from the International Air Transport Association (IATA).

Keywords: time series; anomaly detection; unsupervised learning; kernel density estimation; missing data

1. Introduction

Being able to detect anomalies has many applications, including in the fields of medicine and healthcare management [1,2]; in data acquisition, such as filtering out anomalous readings [3]; in computer security [4]; in media monitoring [5]; and many in the realm of public safety such as identifying thermal anomalies that may precede earthquakes [6], identifying potential safety issues in bridges over time [7], detecting anomalous conditions for trains [8], system level anomaly detection among different air fleets [9], and identifying which conditions pose increased risk in aviation [10]. Given a dataset, anomaly detection is about identifying individual data that are quantitatively different from the majority of other members of the dataset. Anomalous data can come in a variety of forms such as an abnormal sequence of medical events [11] and finding aberrant trajectories of pantograph-caternary systems [12]. In our context, we look for time series of aviation safety incident frequencies for fleets of aircrafts that differ substantially from the rest. By identifying the aircraft types or airports that have significant different patterns of frequencies of specific incidents, our model can provide insights on the potential risk profile for each aircraft type or airport and highlight areas of focus for human analysts to perform further investigations.

Identifying anomalous time series can be divided into different types of anomalous behaviour [13] such as: point anomalies (a single reading is off), collective anomalies (a portion of a time series that reflects an abnormality), or contextual anomalies (when a time series behaves very differently from most others). Identifying anomalous time series from a collection of time series, as in our problem, can be done through dimensionality reduction (choosing representative statistics of the series, applying PCA, and identifying points that are distant from the rest) and through studying dissimilarity between curves (a variant of classical clustering like kmeans) [14]. After reducing the dimension, some authors have used entropy-based methods, instead, to detect anomalies [15]. Archetypoid analysis [16] is another method, which selects time series as archetypeoids for the dataset and identifies anomalies as those not well represented by the archetypeoids. Very recently, authors have used a generalization of Isolation Forests to identify anomalies [17] and have examined the Fourier spectrum of time series and looked at shifting frequency anomalies [18]. Our approach, like Functional Isolation Forest, is geometric in flavor and we employ Kernel Density Estimation and analysis of Fourier modes to detect anomalies.

In this manuscript, we present two alternative means of anomaly detection based on Kernel Density Estimation (KDE) [19]. We use two approaches: the first and simplest considers each time series as element of a Hilbert space $\mathcal{H}$ and employs KDE, treating each time series in $\mathcal{H}$ as if it were a point in one-dimensional Euclidean space, placing a Gaussian kernel at each curve with scale parameter $\xi > 0$. We refer to this as the **point approach** to Functional KDE Anomaly Detection, because each curve in $\mathcal{H}$ is treated as a point. This approach then *formally* generates a proxy for the "probability density" over $\mathcal{H}$. Anomalous series are associated with smaller values of this density. This is distinct from considering a single time series as collection of points sampled from a distribution and using KDE upon points in the time series as has been done before [20]. This is a very simple, and seemingly effective method, with ξ chosen as a hyper-parameter. We also present a **Fourier approach**, which approximates a probability density over $\mathcal{H}$ through estimating empirical distributions for each Fourier mode with KDE. This allows us to estimate the likelihood of a given curve. Curves with lower likelihoods are more anomalous. Both methods naturally handle missing data, without interpolating. In real flight operations, sometimes it is not possible to capture and record complete information because incident data is documented from voluntary reporting, which may result in incomplete datasets. Therefore, model robustness to the impact of missing data is crucial to derive the correct understanding, which may save human lives and prevent damaged aircrafts.

The rest of our paper is organized as follows: in Section 2, we present the details and implementation of our methods; in Section 3, we conduct some experiments to investigate the strengths and weaknesses of the approaches and compare them with three other methods (Functional Isolation Forest available in `Python` and the PCA and functional boxplot methods available in `R`); following this, we apply our techniques to data from the International Air Transport Association (IATA); finally, in Section 4, we discuss our results and present some recommendations.

2. Functional Kernel Density Estimation

2.1. Review of Kernel Density Estimation

We first recall KDE over $\mathbb{R}^d$, $d \in \mathbb{N}$. Given a sample $\mathcal{S} \subset \mathbb{R}^d$ of n points from a distribution with probability density function (pdf) $f : \mathbb{R}^d \to [0, \infty)$ with $\int_{\mathbb{R}^d} f(x)\mathrm{d}x = 1$, KDE provides an empirical estimate for the probability density given by [19]

$$\tilde{f}(x) = \frac{1}{n} \sum_{y \in \mathcal{S}} |\Xi|^{-1/2} K(\Xi^{-1/2}(x - y)) \tag{1}$$

where Ξ is a symmetric, positive definite matrix known as the *bandwidth matrix* and K is a Kernel function. We choose the form of a multivariate Gaussian function so

$$\tilde{f}(x) = \frac{1}{n} \sum_{y \in \mathcal{S}} \frac{e^{-\frac{1}{2}(x-y)^T \Xi^{-1}(x-y)}}{(2\pi)^{d/2}|\Xi|^{1/2}}, \qquad x \in \mathbb{R}^d \tag{2}$$

and we choose [19]

$$\Xi = \alpha \operatorname{diag}(\tilde{\sigma}_1, \tilde{\sigma}_2, ..., \tilde{\sigma}_d) \tag{3}$$

where $\tilde{\sigma}_i$ is the sample standard deviation of the ith-coordinate of the sample points in $\mathcal{S}$ and

$$\alpha = \left(\frac{4}{(d+2)n}\right)^{1/(d+4)}. \tag{4}$$

We used tilde (~) rather than hat (ˆ) to denote estimators as later on we use the hats for Fourier Transform modes and wish to avoid ambiguities. In general tildes will be used for estimates derived from samples.

2.2. Setup, Assumptions, and Notation

The proposed methods are applicable to situations where we look for anomalous time-series relative to the sample we have. We study time series, which we consider more abstractly as being discrete samples from curves of form $x(t)$ where $x : [0, T] \to \mathbb{R}$ for some $T > 0$. The space of all such curves is quite general and we limit the scope to Hilbert spaces on $[0, T]$. For example, we may consider spaces $\mathcal{H} = L^2([0, T])$ or $H^1([0, T])$, the space of square integrable functions or the space of square integrable functions whose derivative is also square integrable, respectively. Within our Hilbert space, $\mathcal{H}$, there is an inner product $(\cdot, \cdot) : \mathcal{H}^2 \to \mathbb{C}$ and an induced norm, $||\cdot|| : \mathcal{H} \to [0, \infty)$ where $||x|| = (x, x)^{1/2}$. With this norm, we can define distances between elements of $\mathcal{H}$.

Observations are made at p different times, $t_0, t_1, ..., t_{p-1}$ where $t_i = i\Delta$ with $\Delta = T/p$ and $i = 0, 1, ..., p - 1$. We also have $t_p = T$, but this time is not included in the data. Although observations are made at these times, some time series could have missing values. When a value is missing, we will say its "value" is Not-a-Number (NaN). While the set of observation points are uniformly spaced, the times at which a given time series has non-NaN values may not be.

We denote by n the number of time series observed, given to us as a sample of form $\mathcal{X} = \{\{(t_j^{(k)}, x_j^{(k)})\}_{j=0}^{P_k-1}\}_{k=1}^n$, where $k = 1, ..., n$ indexes the time series, P_k is the number of available (i.e., non-NaN) points for time series k, $0 \le t_0^{(k)} < t_1^{(k)} < ... < t_{P_k-1}^{(k)} < T$ are the times for series k, with corresponding non-NaN values $x_0^{(k)}, x_1^{(k)}, ..., x_{P_k-1}^{(k)} \in \mathbb{R}$.

2.3. Preprocessing

The methods often performed better if we normalized the data by a standard centering and rescaling. At each fixed observation time, the values of the time series were shifted to have mean zero and then rescaled to have unit variance. When the variance was already zero, the values were mapped to 0. Further remarks are given in Section 4.

Even though our methods do not assume stationary or other similar properties, applying transformations to the data before applying them can be done. For example, we may wish to make the series stationary, or to extract some characteristics (e.g., the cyclical part, or the seasonal part). This can be useful if we want to focus on finding specific types of anomalies.

2.4. Point Approach to Functional KDE Anomaly Detection

Our first method can be summarized as follows: treat each $x \in \mathcal{H}$ as a point in one dimension. Select a value for the KDE scale hyper-parameter $\xi > 0$, and define a score functional over $\mathcal{H}$ by

$$S_P[a] = \sum_{x \in \mathcal{X}} e^{-\frac{||x-a||^2}{2\xi^2}}, \quad a \in \mathcal{H}, \tag{5}$$

which, at least formally, can be thought of as a proxy to a "probability density" functional. More rigorously, one should consider measures on Hilbert spaces [21]. Assuming anomalous curves are truly rare, they should be very distant from the majority of curves and $S_P[\cdot]$ should be smaller at such curves. See Figure 1 for a conceptual illustration. We find that choosing ξ to be the mean of $\{||a||\}_{a \in \mathcal{X}}$ to work well; another natural choice would be the median. These choices are natural because they represent a natural size/scale for the series. This approach can also be interpreted from a Fourier perspective which we remark on in Appendix A.
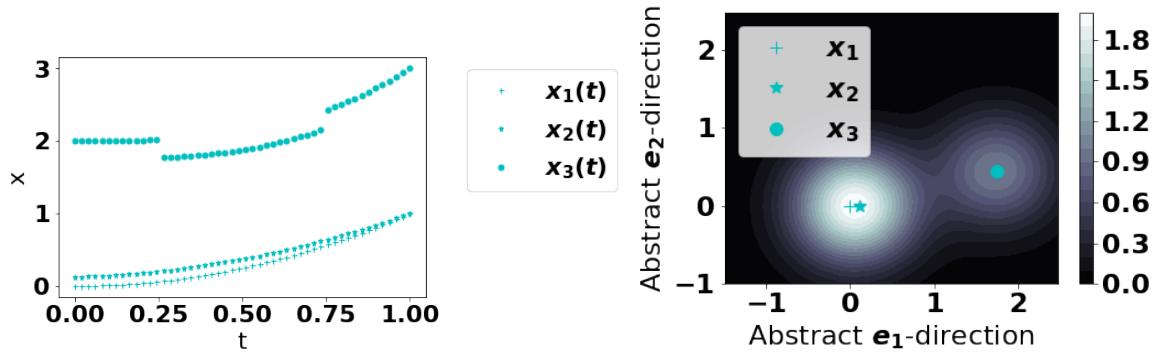

Figure 1. A visual depiction of the Point method. The curves are time series in a Hilbert space $\mathcal{H}$ but after applying KDE, there is a score associated to each point in $\mathcal{H}$. In the cartoon, curves 1 and 2 are similar and curve 3 is anomalous. (**Left**): the time series. (**Right**): a representation of them with associated scores in the color scale. In reality, the space is infinite dimensional and this is only a conceptual illustration.

This method can be implemented with the following steps:

1. Choose $\xi > 0$.
2. For each $x \in \mathcal{X}$, compute its score from (5) where, for example, in the case of $\mathcal{H} = L^2([0, T])$,

$$||x - a||^2 = \int_0^T |(x(t) - a(t)|^2 \mathrm{d}t. \tag{6}$$

3. Identify anomalies as curves with the lowest score.

The integral in (6), even with some data points missing, can be computed as below:

1. To compute $I = \int_0^T |(x(t) - a(t)|^2 \mathrm{d}t$, determine all t-values where both x and a are not NaN. Call these $t_0^*, t_1^*, ..., t_{r-1}^*$.
2. Define $t_r^* = T - t_{r-1}^* + t_0^*$, $x_r^* = x_0^*$ and $y_r^* = y_0^*$.
3. Estimate the integral as

$$I \approx \frac{1}{2} \sum_{m=0}^{r-1} (t_{m+1}^* - t_m^*)(|x(t_m^*) - y(t_m^*)|^2 + |x(t_{m+1}^*) - y(t_{m+1}^*)|^2).$$

This is a second-order accurate (trapezoidal) approximation to I where we have extended the signal periodically at the endpoint. This ensures that in a pathological case such as there being only a single point of observation for the integrand with value v, then the inner product evaluates to Tv.

2.5. Fourier Approach to Functional KDE Anomaly Detection

We first observe that most Hilbert spaces of interest such as $L^2([0,T])$ have a countable, orthogonal basis $\mathcal{B} = \{\exp(2\pi ikt/T) | k \in \mathbb{Z}\}$. By considering time series as being represented by these basis vectors, we can more accurately consider a true probability density over $\mathcal{H}$. In practice, we pick $L \in \mathbb{N}$ large and represent $a \in \mathcal{H}$ by

$$a(t) \approx \sum_{j=-L}^{L} \hat{a}_k \mathrm{e}^{2\pi ikt/T}.$$

Then, up to a Fourier mode of size L, we can define a probability density at $a \in \mathcal{H}$ by $\prod_{k=-L}^{L} \zeta_k(\hat{a}_k)$ where ζ_k is a pdf over $\mathbb{C}$ for mode k.

Our time series are discrete with finitely many points so we consider a Non-Uniform Discrete Fourier Transform (NUDFT). To estimate the probability density over $\mathcal{H}$ at a, we:

1. Compute $p^* = \min\{P_1, P_2, ..., P_n\}$.
2. Compute the Discrete Fourier coefficients

$$\hat{x}_j^{(k)} = \frac{1}{P_k} \sum_{m=0}^{P_k-1} \exp(-2\pi ijt_m/T)x^{(k)}(t_m)$$

for each $k = 1, ..., n$ and for $j = 0, 1, ..., p^* - 1$.
3. For each $0 \leq j \leq p^* - 1$, use KDE to estimate the pdf over $\mathbb{C}$ for $\hat{x}_j$, by using KDE (Equations (2)–(4)) for $\mathbb{R}$ or $\mathbb{R}^2$ when the coefficients are all purely real/imaginary or contain a mix of real and imaginary components, respectively. Call the empirical distribution $\tilde{\zeta}_j$ for each j.
4. For any $a \in \mathcal{H}$ define an estimated pdf via

$$\rho_F[a] = \prod_{j=0}^{p^*-1} \zeta_j(\hat{a}_j). \tag{7}$$

5. Let the score of $a \in \mathcal{H}$ be
$$S_F[a] = \log \rho_F[a]. \tag{8}$$

6. Identify anomalies in $\mathcal{X}$ as those whose scores given by (8) are smallest.

Due to missing data, this method does lose some information since the higher Fourier modes necessary to fully reconstruct a given time series may be discarded. Additionally, as the missing data may result in non-uniform sampling, the typical aliasing of the Discrete Fourier Transform does not take effect. In general for one of the series $x^{(k)}$, we will not have $\hat{x}_{P_k-j}^{(k)} = \overline{\hat{x}_j^{(k)}}$, where the bar denotes complex conjugation. See the remark on aliasing in Appendix B.

In multiplying the pdfs in each mode to estimate the probability density at a point in the Hilbert space, we have implicitly assumed the modes are independent. It may seem intuitive to decouple the modes by applying a Mahalanobis transformation upon the modes prior to KDE, but this results in poor outcomes. Thus, this implicit independence seems to work well in practice, without adjustments.

A Discrete Fourier transform of a signal $x_0, x_1, ..., x_{P_S-1}$ measured at times $\tilde{t}_0, \tilde{t}_1, ..., \tilde{t}_{P_S-1}$ is a representation in a new basis $\{e^{(k)}\}_{k=0}^{P_S-1}$ where $e_j^{(k)} = \mathrm{e}^{2\pi ik\tilde{t}_j/T}$ for $j = 0, ..., P_S - 1$. In general, such a basis of vectors for a NUDFT will not be orthogonal [22]. However, if $m = p - P_S \ll p$ and the $\tilde{t}$'s are a subset of a uniformly spaced set of times, we can show that the vectors are *almost* orthogonal with a cosine similarity of size $O(m/p)$. Details appear in Appendix C. This orthogonality is not strictly necessary to run the method, but doing so allows a deeper justification of multiplying the pdfs in each mode if the Fourier modes are truly independent because the Discrete Fourier Transform is then approximately a projection onto an orthogonal basis of modes, each of which are independent.

3. Method Performance

We begin by illustrating the performance of our methods for some synthetic data and compare Functional KDE to other methods. The first one is the Functional Boxplot (FB) [23]. The `fbplot` function in the R package `fda` is used to obtain a center outward ordering of the time series based on the band depth concept which is a generalization to functional data of the univariate data depth concept [24]. The idea is that anomalous curves will be the ones with the largest ranks, that is, the ones that are farther away from the center. The second method is the recently proposed Functional Isolation Forest (FIF) [17], which is also depth-based and assigns a score to a curve, with higher values indicating that it is more anomalous. We used the code provided for FIF directly on GitHub [25] with the default settings given. The third is the method proposed in [26] and implemented in the R package `anomalousACM` [27]. This method works in three steps: (i) extract features (e.g., mean, variance, trend) from the time-series; (ii) use Principal Component Analysis (PCA) to identify patterns; (iii) Use a two dimensional outlier detection algorithm with the first two principal components as inputs. It will be referenced as the PCA method in what follows After testing them on synthetic data, we apply our techniques to real data to identify anomalies in time series for aviation events.

The methods against which we compare our methods did not have standard means of managing missing entries. For these methods, we replace missing data (NaN) in a series using `Python`'s default interpolation scheme. For the methods proposed in this paper, we do not have to use imputation.

3.1. Synthetic Data

We apply the Point and Fourier Approaches to Functional KDE, Functional Boxplot, and Functional Isolation Forest to the two scenarios described below.

Scenario 1 : we define a base curve

$$x_0(t) = a_0(1 + \tanh(b_0(t - t_0))) + c_0 \sin(\omega_0 t/T),$$

with $a_0 = 5$, $b_0 = 2$, $T = 50$, $t_0 = T/2 = 25$, and $\omega_0 = 2\pi$. Ordinary curves are generated via

$$x_0(t) + \epsilon(t),$$

where $\epsilon(t)$ represents Gaussian white noise at every t with mean $\mu_g = 0$ and standard deviation $\sigma_g = 0.05$. We then consider a series of 7 anomalous curves:

- $C_1(t) = x_0(t)\left(1 + r_1 \frac{(t-t^*)^2}{1+(t-t_0)^2}\Theta(t - t_0)\right) + \epsilon(t)$, where $r_1 = 0.05$ and Θ denotes the Heaviside function. Thus, the function is scaled up after t_0.
- $C_2(t) = x_0(t) + \left(1 + r_2\Theta(t - t_0)\right)\epsilon(t)$, where $r_2 = 3$. Thus, the noise is larger after t_0.
- $C_3(t) = x_0(t) - r_3(t - t_0)\Theta(t - t_0) + \epsilon(t)$, where $r_3 = 0.05$. Thus, there is a decreasing component added to the function after t_0.
- $C_4(t) = 2a_0\Theta(t - t_0) + c_0 \sin(\omega_0 t/T) + \epsilon(t)$, i.e., the tanh is replaced by a discontinuous function.
- $C_5(t) = x_0(t) + \mathcal{E}(t)$, where $\mathcal{E}(t)$ represents an exponential random variable at every t with mean 0.05.
- $C_6(t) = a_0(1 + \tanh(2b_0(t - t_0))) + c_0 \sin(\omega_0 t/T) + \epsilon(t)$, which has a slightly steeper transition rate than the base curve.
- $C_7(t) = a_0(1 + \tanh(b_0(t - t_0)) + c_0 \sin((1 + r_7 t/T)\omega_0 t/T) + \epsilon(t)$, where $r_7 = 0.1$ so the frequency increases with time.

Over 50 trials, we generate 70 time series, 63 normal curves, and 7 anomalous curves with each of C_1 through C_7 being used once. See Figure 2 for an illustration. We used a uniform mesh with 50 points, $0, 1, ..., 49$. Since we used a 9 : 1 ratio of regular to anomalous series, successful methods, after ranking curves in ascending order of "regular," should rank anomalous curves as among the bottom 10%. We can also determine the 95th percentile for the percentile rank of each curve, to give an estimate for

how much of the data would need to be re-examined to capture such anomalies. These trials can also be done by dropping data points independently at random with a fixed probability to simulate missing data. We ran sets of trials with 0% and 10% of drop probabilities. Results for the mean percentile rank and 95th percentile of the percentile ranks are presented in Tables 1 and 2.
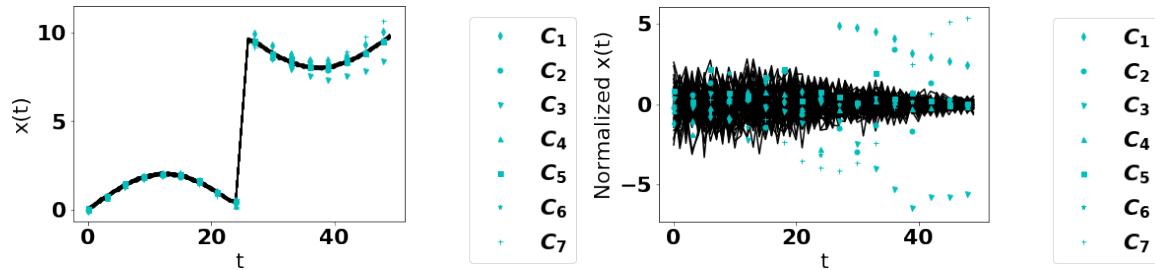

Figure 2. Plot of 63 normal curves and the 7 anomalous curves $C_i(t)$, $i = 1, ..., 7$. Left: un-normalized. Right: normalized.

Table 1. Mean percentiles (out of 100) for curves C_1–C_7 in Scenario 1. A correct classification is a percentile less than or equal to 10 (in bold in the table). The *-N* suffix denotes the data were normalized by the pre-processing described in Section 2.3; the *-U* suffixed denotes the data were un-normalized. Note that method FB is not affected by the normalization.

Method	Lost	C_1	C_2	C_3	C_4	C_5	C_6	C_7
Point-N	0%	**4.3**	**5.8**	**1.4**	21	12	24	**2.9**
Point-U	0%	**4.3**	**5.7**	**1.4**	14	**8.8**	16	**2.9**
Fourier-N	0%	**4.7**	**1.9**	**2.8**	28	51	29	**8.5**
Fourier-U	0%	**5.8**	**4.0**	**2.8**	38	43	35	**1.7**
FIF-N	0%	51	24	72	56	79	58	13
FIF-U	0%	19	**2.2**	**4.8**	18	53	19	**5.3**
PCA	0%	**7.5**	20.	**4.3**	**7.5**	53	**9.4**	11
FB	0%	**4.5**	**5.6**	**1.8**	36	21	37	**2.5**
Point-N	10%	**4.3**	**6.0**	**1.4**	23	18	29	**2.9**
Point-U	10%	**4.3**	**5.7**	**1.4**	20.	14	23	**2.9**
Fourier-N	10%	**4.3**	**4.5**	**2.5**	28	43	36	**4.0**
Fourier-U	10%	45	59	50	46	49	53	49
FIF-N	10%	50.	21	75	48	74	51	13
FIF-U	10%	45	15	29	42	48	44	32
PCA	10%	32	20.	**6.1**	36	47	35	**7.7**
FB	10%	**7.5**	**8.7**	**4.2**	47	24	49	**5.1**

Scenario 2: we utilized the testing examples of Staerman et al. [17]. The data consist of 105 time series over $[0, 1]$ with 100 time points. There are 100 regular curves defined by $x(t) = 30(1 - t)^q t^q$ where q is equi-spaced in $[1, 1.4]$–thus there is a large family of normal curves. Then, there are 5 anomalous curves:

- $D_1(t) = 30(1 - t)^{1.2} t^{1.2} + \beta \chi_{[0.2, 0.8]}$, where β is chosen from a Normal distribution with mean 0 and standard deviation 0.3 and χ_I is the characteristic function of I (there is a jump discontinuity at 0.2 and 0.8).
- $D_2(t) = 30(1 - t)^{1.6} t^{1.6}$, being anomalous in its magnitude.
- $D_3(t) = 30(1 - t)^{1.2} t^{1.2} + \sin(2\pi t)$.
- $D_4(t) = 30(1 - t)^{1.2} t^{1.2} + 2\chi_{\{\tau\}}$, where $\tau = 0.7$ is a single point.
- $D_5(t) = 30(1 - t)^{1.2} t^{1.2} + \frac{1}{2} \sin(10\pi t)$.

Each curve was sampled uniformly at 100 points. We did not drop any data points and, owing to the limited randomness, we only present the results of one trial. The results are presented in Table 3.

Table 2. The 95th percentile of the percentile ranks (out of 100) for curves C_1–C_7 in Scenario 1. See Table 1 caption for *-N* vs. *-U* distinction.

Method	Lost	C_1	C_2	C_3	C_4	C_5	C_6	C_7
Point-N	0%	4.3	6.5	1.4	47	42	62	2.9
Point-U	0%	4.3	5.7	1.4	30.	12	46	2.9
Fourier-N	0%	8.6	3.6	4.3	67	99	74	13
Fourier-U	0%	5.7	4.3	4.2	84	94	77	2.9
FIF-N	0%	84	69	92	90	100	96	30
FIF-U	0%	57	7.5	11	55	97	53	11
PCA	0%	25	67	7.1	17	96	27	47
FB	0%	5.7	5.7	2.9	75	74	75	2.9
Point-N	10%	4.3	7.1	1.4	52	51	72	2.9
Point-U	10%	4.3	5.7	1.4	46	43	59	2.9
Fourier-N	10%	6.5	10.	5.1	61	91	89	5.7
Fourier-U	10%	84	97	92	93	93	92	89
FIF-N	10%	81	67	97	91	99	95	39
FIF-U	10%	84	27	56	88	95	94	56
PCA	10%	82	60.	31	73	94	73	19
FB	10%	11	13	10.	75	75	75	9.4

Table 3. Percentiles (out of 100) for curves D_1–D_5 in Scenario 2. A correct classification is a percentile less than or equal to 4.8 (in bold in the table) since $5/105 = 4.8\%$. See Table 1 caption for *-N* vs. *-U* distinction.

Method	D_1	D_2	D_3	D_4	D_5
Point-N	74	**0.95**	6.7	73	**1.9**
Point-U	83	**0.95**	44	85	71
Fourier-N	**3.8**	**4.8**	**1.9**	**2.9**	**0.95**
Fourier-U	**1.9**	8.6	30	**0.95**	**2.9**
FIF-N	**1.9**	28	**3.8**	10	**0.95**
FIF-U	**1.9**	**2.9**	**3.8**	**4.8**	**0.95**
PCA	**4.8**	**2.9**	**3.8**	**1.9**	**0.95**
FB	75	**0.95**	21	75	75

Our method is unsupervised and thus the distinction as to what constitutes an anomaly requires considering a curve's score relative to the others and making a decision based upon this. This can involve human judgment. However, since our method returns a scalar score, we can also use a univariate outlier test on the score to formally test the hypothesis H_0: There are no anomalies. The Rosner test [28] is such a test and is available in the `R` package `EnvStats` [29]. In Appendix D, by considering scenarios where anomalies are present or absent, we show the validity of this approach.

3.2. Aviation Safety Reports

We now consider how our methods behave in identifying anomalous time series for aviation safety events. A discussion on method performance is deferred to Section 4.

We were provided IATA data for safety-related events of different types on a month-by-month basis from 2018–2020 for different aircraft types and airports. Aircraft types were given IDs from 1 to 64 (not every ID in the range was included). We were also given separate data pertaining to flight frequency in order to normalize and obtain event rates (cases per 1000 flights). Events of interest could include phenomena occurring during a flight such as turbulence or striking birds, or physical problems such as a damaged engine. We study two events: A and B. Event Type A is a contributing factor for one specific type of accident; Event Type B is the aircraft defense against that type of accident. To illustrate our method while preserving the confidentiality of the data, we do not state what A and B represent.

We plot histograms for the scores of Type A and B Events in Figure 3. These histograms suggest that, for events A and B, anomalous curves could be those with scores below 10 for the Point approach.

Then, we consider curves anomalous by the Fourier method if they have scores below -60 for event A and -30 for event B. As the method is unsupervised, the notion of where to draw the line of being anomalous is somewhat subjective. The idea here is to raise a flag so that experts can investigate the anomalous cases more closely. The aircraft types identified as anomalous for both methods are presented in Table 4. It appears for these data, the curves deemed anomalous by the Fourier method are a subset of the curves deemed anomalous by the Point method. In Figure 4, we plot the anomalous curves (with markers) along with the normal curves (dotted lines) for the fleet IDs that were common to both approaches.
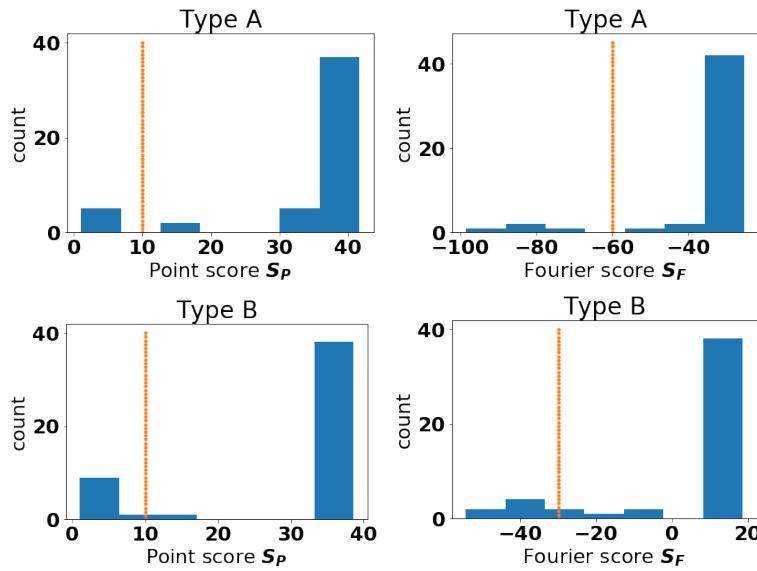
Figure 3. Histogram of scores for Point and Fourier methods for Type A Point (**top-left**), Type A Fourier(**top-right**), Type B Point (**bottom-left**) and Type B Fourier (**bottom-right**). The dashed vertical line represents the division we chose between anomalous (left of line) and normal (right of line). The Sturges estimate was used to set bin widths [30].
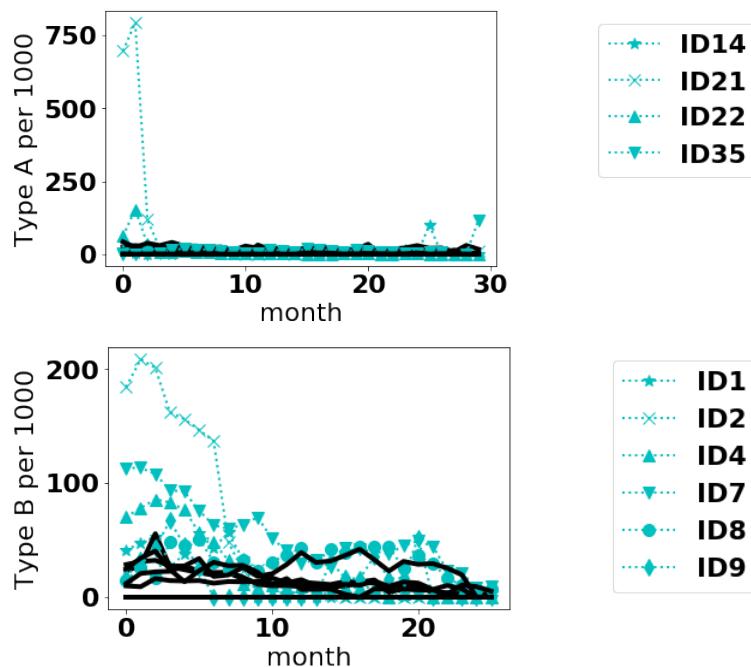
Figure 4. Plots of the time series for Type A and Type B events. Anomalous are dotted curves with markers in the legend; normal curves are solid black curves.

Table 4. IDs of anomalous flights for events A and B. Columnwise, the bolded IDs are common to both methods for a given event type.

Method	Type A	Type B
Point	**14, 21, 22**, 23, **35**	**1, 2, 4**, 5, 6, **7, 8, 9**, 25, 34
Fourier	**14, 21, 22, 35**	**1, 2, 4, 7, 8, 9**

4. Discussion and Conclusions

4.1. Method Performance

From Tables 1 and 2 with regards to Scenario 1, the Point method and FB are superior. They correctly classify C_1–C_3 and C_7 as anomalous. The Point and PCA methods significantly outperform the other methods in the more difficult C_4–C_6 curves. With these data, the Point method generally performs better without normalization. Generally all methods failed to identify the replacement of Gaussian white noise with exponential noise (C_5) as anomalous, although the un-normalized Point approach succeeded. Additionally, all methods considered, except PCA, had difficulty identifying the discontinuous replacement of the hyperbolic tangent (C_4) and a slightly steeper hyperbolic tangent (C_6). The PCA method was not as good as the others for identifying the noise increases (C_2) and the frequency increase with time (C_7). This suggests that not all methods are effective at detecting the same types of anomalies and that they may be complementary.

From Table 3 for Scenario 2, the Fourier approach with normalized data, FIF on un-normalized data, and PCA classify equally correctly. Data can always be normalized and this is therefore not a problem for the Fourier method. In this example, the Point approach fares better with normalization. However, this method and the FB method are not as effective as the FIF, PCA, and Fourier methods.

From our experiments, when there was a large family of curves as with Scenario 2, the Fourier method performed better at detecting anomalies, especially when provided normalized data. But when the family of curves were all close to the same, except for noise, the Point method was better, with or without normalization. Providing more theoretical understanding as to whether these are general phenomena is left for future work.

4.2. Aviation Safety Data

From Figure 4, it appears the methods can detect different sorts of anomalies. In the case of Type A events, the anomalous curves appear to have anomalously large values at an isolated point or over small range of values. The anomalies in Type B events are more interesting and subtle. Even some of the normal curves have sizeable event frequencies, sometimes even exceeding the anomalous curves. But on the average it seems the anomalous curves are higher. In the case of curve 9, the reason it is deemed anomalous is not immediately intuitive. Whether such differences are of a concern to safety would require follow-up from safety inspectors.

To prevent aviation accidents, identifying the potential hazards and risks before they evolve into accidents is the key to proactive safety management. While collecting and analyzing data manually is a time-consuming process, especially on a global scale, the risk identification process may remain reactive process if there is not an automated process. The application of the anomaly detection will enable proactive data-driven risk identification in global aviation safety, by continuously monitoring aviation safety data across multiple criteria (e.g., airport, aircraft type and date), then automatically raising a flag when the model detects any anomalous patterns.

The proposed model shows potential value in automatically detecting potential risk areas with robustness from missing data; however, the interpretation of the model still requires future study. As safety risk is an outcome of complex interactions between multiple factors, including human, machine, environment, and other hidden factors, understanding the full context of such risk requires in-depth investigation and validation from multiple experts. While the model can identify some

anomalous patterns, this does not take into consideration the interactions. For example, some aircraft fleets fly more frequently over certain pathways than others. Thus, some differences identified as anomalous due to aircraft type may actually stem from location. Therefore, there will always be a human layer between the model and the interpretation of the model.

4.3. Comments on the Models

There are various degrees of freedom the proposed methods allow for, which are worth noting. Firstly, the point method could be generalized to compute $H^1([0, T])$, and higher Sobolev norms too, but that could lead to additional hyper-parameters in how heavily to weigh the derivative terms. With the Fourier approach, it may seem more appropriate to replace the NUDFT with a weighted combination of terms that more accurately reflects the non-uniform spacing, i.e., a Riemann Sum. Interestingly such an approach tends to make the results slightly worse, hence our choice to use the standard NUDFT.

We anticipate these methods perform well when the time series are sampled at regular intervals and a small portion of entries are missing. If the number of missing entries is very large, this makes inner products computed with the Point method less accurate (without additional interpolations) and the preprocessing of shifting and rescaling could result in poorer outcomes due to a limited sample size upon which to base the normalizations. For many applications, however, most data are present.

4.4. Future Work

We note that our proposed methods aim to identify anomalous time series relative to the sample taken. In general, even if all time series are sampled from the same distribution, due to low probability events, some time series could still be anomalous relative to the sample given. As such, our work has mostly been an empirical investigation of the methods; however, by adding further assumptions on the underlying distribution of time series, it could be possible to obtain a more theoretical basis for the method performance. This would be worth investigating, but is beyond our current work. Going hand-in-hand with this theory it would be interesting to investigate the optimal choice of ξ in the point approach, to understand how the Fourier modes being treated as independent works as effectively as it does, or to more rigorously establish classes of problems when the Point or Fourier approaches are superior.

In conclusion, we have presented two approaches to detecting anomalous time series using KDE to generate functionals to score a series for its degree of anomalousness. The methods handle missing data and perform well in comparison to other methods.

Appendix A. Fourier Perspective of the Point Method

From Parseval's identity [31], we can also write the terms of (5) as

$$e^{-\frac{||x-a||^2}{2\zeta^2}} = e^{-\frac{\sum_{k\in\mathbb{Z}}|\hat{x}_k-\hat{a}_k|^2}{2T\zeta^2}},$$

i.e., each kernel can be thought of as a Gaussian in $\mathbb{C}^\infty$ with constant variance in all directions. Unfortunately this thinking can only be true in spirit because such a series would not be in $L^2([0,T])$ as $\sum_{k\in\mathbb{Z}}|\hat{a}_k|^2$ would almost surely diverge.

Appendix B. Aliasing

Observe that if the *real* time series $x_0, x_1, ..., x_{N-1}$ is observed at p equally spaced points $t_j = j\Delta$, $\Delta = T/p$, $j = 0, ..., p-1$ then for $j = 1, ..., p-1$:

$$\hat{x}_{p-j} = \sum_{m=0}^{p-1} e^{-2\pi i t_{p-j} m\Delta/T} x_m$$

$$= \sum_{m=0}^{p-1} e^{-2\pi i(p-j)m\Delta/T} x_m$$

$$= \sum_{m=0}^{p-1} e^{2\pi i jm\Delta/T} x_m$$

$$= \overline{\hat{x}_j}$$

where in getting from the first to second line we used $\exp(-2\pi i pm\Delta/T) = \exp(-2\pi im) = 1$. On the other hand, if data are only observed at $\tilde{t}_0 < \tilde{t}_1 < ... < \tilde{t}_{P_S}$, a subset of the times $t_0, ..., t_p$ with $P_S < p$ then $\tilde{t}_j$ is not, in general $j\Delta$ and the identity does not hold.

Appendix C. Approximate Orthogonality

Before our approximate orthogonality result, we first define the standard inner product for vectors over $\mathbb{C}^N$:

$$(x, y) = \sum_{j=1}^{N} \bar{x}_i y_i.$$

Theorem A1 (Approximate Orthogonality). *Let $t_j = j\Delta$ for $j = 0, 1, ..., p-1$ where $\Delta = T/p$ for $T > 0$. Let $\{\tilde{t}_j\}_{j=0}^{P_S-1} \subset \{t_0, t_1, ..., t_{p-1}\}$. Define $m = p - P_S$ and define the basis vectors $\{e^{(k)} = e^{2\pi i k\tilde{t}_j/T}, j = 0, ..., P_S | k = 0, ..., P_S\}$. Then*

$$\frac{(e^{(k)}, e^{(k')})}{|e^{(k)}||e^{(k')}|} = \begin{cases} 1, & k = k' \\ O(m/p), & k \neq k' \end{cases}.$$

In other words the cosine similarity of the two vectors is either 1 or $O(m/p)$.

Proof. We trivially note that $|e^{(k)}| = \sqrt{P_S}$ for any k. Next, if $k = k'$ then

$$\frac{(e^{(k)}, e^{(k')})}{|e^{(k)}||e^{(k')}|} = \frac{1}{P_S} \sum_{j=0}^{P_S-1} 1$$

$$= 1.$$

Let us define the set $B = \{j|t_j \notin \{\tilde{t}_k | k = 0, ..., P_S\}$ for $j = 0, ..., p\}$, i.e., it is a listing of all regular time

values that have been lost in only observing at the $\tilde{t}$'s. Note that $|B| = m$. Also let $q = k' - k$ so that when $k \neq k'$:

$$
\frac{(e^{(k)}, e^{(k')})}{|e^{(k)}||e^{(k')}|} = \frac{1}{P_S} \sum_{j=0}^{N_S-1} \exp(2\pi i q \tilde{t}_j / T)
$$

$$
= \frac{1}{P_S} \left(\sum_{j=0}^{p-1} \exp(2\pi i q t_j / T) - \sum_{j \in B} \exp(2\pi i q t_j / T) \right)
$$

$$
= \frac{-1}{P_S} \sum_{j \in B} \exp(2\pi i q t_j / T)
$$

where in arriving at the final equality we used that the sum of $e^{2\pi i q t_j}$ over $j = 0, ..., p - 1$ is 0 ($1 + \eta + \eta^2 + ... + \eta^{p-1} = 0$ if $\eta^p = 1$ and $\eta \neq 1$). As each term in the remaining sum is bounded by 1 and $|B| = m$, we have that

$$
\frac{(e^{(k)}, e^{(k')})}{|e^{(k)}||e^{(k')}|} = O(m/P_S)
$$

$$
= O(m/p).
$$

□

Appendix D. Rosner Test on the Scores

We briefly explore the capacity of our methods to test the hypothesis H_0: There are no anomalies. The idea is to compute the scores from our methods, which are scalars, and use a univariate outlier test on them. In this experiment, we use Rosner test [28] with the R package `EnvStats` [29]. Specifically, we test H_0 for each of the 50 trial datasets of Scenario 1 and report the proportion of time the null hypothesis is rejected at the $\alpha = 0.05$ level. In this case, the proportion of rejection measures the power of the test and we wish to have the highest values possible. However, to verify the validity of the test, we also run the tests on the samples containing only the 63 normal curves. This time, we want the proportion of rejection to be close to the level $\alpha = 0.05$. The results, presented in Table A1, show that this method is working. Setting aside the un-normalized Fourier approach (Fourier-U) with 10% of missing data, the proportion of rejection varies between 0.02 and 0.08 when the data contain no anomalies, showing that the test is able to maintain its prescribed level. When the data contain anomalies, the power ranges between 0.98 and 1, showing that the anomalies are detected in almost all cases. The only exception is the Fourier-U method with 10% of missing data, which never rejects H_0 whether or not the data contain anomalies. But this is consistent with the fact that this method had a very poor performance in this case and was not able to detect the anomalies as seen in Table 1.

Table A1. Proportions of time the null hypothesis H_0 is rejected at the $\alpha = 0.05$ level. See Table 1 caption of main manuscript for -*N* vs -*U* distinction.

Method	Lost	Anomalies Present	H_0 Reject Fraction
Point-N	0%	No	0.04
Point-N	0%	Yes	1
Point-U	0%	No	0.08
Point-N	0%	Yes	1
Fourier-N	0%	No	0.08
Fourier-N	0%	Yes	0.98
Fourier-U	0%	No	0.02

Table A1. *Cont.*

Method	Lost	Anomalies Present	H_0 Reject Fraction
Fourier-U	0%	Yes	1
Point-N	10%	No	0.04
Point-N	10%	Yes	1
Point-U	10%	No	0.08
Point-N	10%	Yes	1
Fourier-N	10%	No	0.06
Fourier-N	10%	Yes	1
Fourier-U	10%	No	0
Fourier-U	10%	Yes	0

References

1. Cleophas, T.J.; Zwinderman, A.H.; Cleophas-Allers, H.I. *Machine Learning in Medicine*; Springer: Berlin/Heidelberg, Germany, 2013; Volume 9.
2. Howedi, A.; Lotfi, A.; Pourabdollah, A. An Entropy-Based Approach for Anomaly Detection in Activities of Daily Living in the Presence of a Visitor. *Entropy* **2020**, *22*, 845. [CrossRef]
3. Chen, X.; Li, Z.; Wang, Y.; Tang, J.; Zhu, W.; Shi, C.; Wu, H. Anomaly detection and cleaning of highway elevation data from google earth using ensemble empirical mode decomposition. *J. Transp. Eng. Part Syst.* **2018**, *144*, 04018015. [CrossRef]
4. Siboni, S.; Cohen, A. Anomaly Detection for Individual Sequences with Applications in Identifying Malicious Tools. *Entropy* **2020**, *22*, 649. [CrossRef]
5. Bernikova, O.; Granichin, O.; Lemberg, D.; Redkin, O.; Volkovich, Z. Entropy-Based Approach for the Detection of Changes in Arabic Newspapers' Content. *Entropy* **2020**, *22*, 441. [CrossRef]
6. Saradjian, M.; Akhoondzadeh, M. Thermal anomalies detection before strong earthquakes (M > 6.0) using interquartile, wavelet and Kalman filter methods. *Nat. Hazards Earth Syst. Sci.* **2011**, *11*, 1099. [CrossRef]
7. Xu, X.; Ren, Y.; Huang, Q.; Fan, Z.Y.; Tong, Z.J.; Chang, W.J.; Liu, B. Anomaly detection for large span bridges during operational phase using structural health monitoring data. *Smart Mater. Struct.* **2020**, *29*, 045029. [CrossRef]
8. Holst, A.; Bohlin, M.; Ekman, J.; Sellin, O.; Lindström, B.; Larsen, S. Statistical anomaly detection for train fleets. *AI Mag.* **2013**, *34*, 33. [CrossRef]
9. Das, S.; Matthews, B.L.; Lawrence, R. Fleet level anomaly detection of aviation safety data. In Proceedings of the 2011 IEEE Conference on Prognostics and Health Management, Montreal, QC, Canada, 20–23 June 2011; pp. 1–10.
10. Ketabdari, M.; Giustozzi, F.; Crispino, M. Sensitivity analysis of influencing factors in probabilistic risk assessment for airports. *Saf. Sci.* **2018**, *107*, 173–187. [CrossRef]
11. Huang, Z.; Lu, X.; Duan, H. Anomaly detection in clinical processes. In *AMIA Annual Symposium Proceedings*; American Medical Informatics Association: Bethesda, MD, USA, 2012; Volume 2012, p. 370.
12. Aydin, I.; Karakose, M.; Akin, E. Anomaly detection using a modified kernel-based tracking in the pantograph–catenary system. *Expert Syst. Appl.* **2015**, *42*, 938–948. [CrossRef]
13. Braei, M.; Wagner, S. Anomaly Detection in Univariate Time-series: A Survey on the State-of-the-Art. *arXiv* **2020**, arXiv:2004.00433.
14. Blázquez-García, A.; Conde, A.; Mori, U.; Lozano, J.A. A review on outlier/anomaly detection in time series data. *arXiv* **2020**, arXiv:2002.04236.
15. Martos, G.; Hernández, N.; Muñoz, A.; Moguerza, J.M. Entropy measures for stochastic processes with applications in functional anomaly detection. *Entropy* **2018**, *20*, 33. [CrossRef]
16. Vinue, G.; Epifanio, I. Robust archetypoids for anomaly detection in big functional data. In *Advances in Data Analysis and Classification*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 1–26.
17. Staerman, G.; Mozharovskyi, P.; Clémençon, S.; d'Alché-Buc, F. Functional Isolation Forest. *Proc. Mach. Learn. Res.* **2019**, *101*, 332–347.

18. Jackson, A.C.; Lacey, S. Seasonality and Anomaly Detection in Rare Data Using the Discrete Fourier Transformation. In Proceedings of the 2019 First International Conference on Digital Data Processing (DDP), London, UK, 15–17 November 2019; pp. 13–17.

19. Hyndman, R.L.; Zhang, X.; King, M.L. Bandwidth selection for multivariate kernel density estimation using mcmc. In Proceedings of the Econometric Society 2004 Australasian Meetings, Melbourne, Australia, 7–9 July 2004; number 120.

20. Guo, Y.; Xu, Q.; Li, P.; Sbert, M.; Yang, Y. Trajectory shape analysis and anomaly detection utilizing information theory tools. *Entropy* **2017**, *19*, 323. [CrossRef]

21. Maniglia, S.; Rhandi, A. Gaussian measures on separable Hilbert spaces and applications. *Quaderni di Matematica* **2004**. Available online: http://siba-ese.unisalento.it/index.php/quadmat/issue/view/775 (accessed on 30 November 2020).

22. Mobli, M.; Hoch, J.C. Nonuniform sampling and non-Fourier signal processing methods in multidimensional NMR. *Prog. Nucl. Magn. Reson. Spectrosc.* **2014**, *83*, 21–41. [CrossRef] [PubMed]

23. Sun, Y.; Genton, M.G. Functional boxplots. *J. Comput. Graph. Stat.* **2011**, *20*, 316–334. [CrossRef]

24. Ramsay, J.O.; Graves, S.; Hooker, G. fda: Functional Data Analysis; R Package Version 5.1.5.1. Available online: https://cran.r-project.org/web/packages/fda/index.html (accessed on 30 November 2020).

25. Staerman, G.; Mozharovskyi, P.; Clémençon, S.; d'Alché-Buc, F. FIF: Functional Isolation Forest. Available online: https://github.com/GuillaumeStaermanML/FIF (accessed on 15 September 2020).

26. Hyndman, R.J.; Wang, E.; Laptev, N. Large-scale unusual time series detection. In Proceedings of the 2015 IEEE international conference on data mining workshop (ICDMW), Atlantic City, NJ, USA, 14–17 November 2015; pp. 1616–1619.

27. Hyndman, R.J.; Wang, E.; Laptev, N. anomalousACM: Unusual Time Series Detection; R Package Version 0.1.0. Available online: https://github.com/robjhyndman/anomalous-acm (accessed on 30 November 2020).

28. Rosner, B. Percentage points for a generalized ESD many-outlier procedure. *Technometrics* **1983**, *25*, 165–172. [CrossRef]

29. Millard, S.P. *EnvStats: An R Package for Environmental Statistics*; Springer: New York, NY, USA, 2013.

30. Sturges, H.A. The choice of a class interval. *J. Am. Stat. Assoc.* **1926**, *21*, 65–66. [CrossRef]

31. Folland, G.B. *Real Analysis: Modern Techniques and Their Applications*; John Wiley & Sons: Hoboken, NJ, USA, 1999; Volume 40.