

Application of wavestrapping statistical technique to estimate an extreme value in train aerodynamics

Antonio R. Andrade¹, Terry Johnson², Julian Stow³

Abstract

This paper explores the application of a statistical analysis technique known as ‘wavestrapping’ to a railway data analysis problem in train aerodynamics. This simulation technique is used to reproduce/reconstruct signals with the same statistical properties of the initial signal(s) and to estimate an extreme value in train aerodynamics. This technique is compared with the method currently used in practice to estimate the peak value of slipstream air speed, according with the European standard EN 14067.

Keywords: railway aerodynamics; simulation; bootstrapping; wavelet analysis; wavestrapping.

1. Introduction

As the railway sector is changing towards a data-driven industry, simulation and statistical methods are becoming more and more important in order to synthesize large amounts of data into performance indicators that can support decision-making processes. Larger amounts of data provide further opportunities in risk analysis and safety assessment, especially in extreme situations where peak values play an important role.

Simulation methods and statistical techniques currently used in risk analysis and safety assessment tend to benefit from Monte Carlo approaches, i.e. by assuming some probability distributions for the inputs and by simulating a sufficiently large number of times the system, in order to estimate an extreme value for the outputs (e.g. 99th percentile). Monte Carlo approaches are particularly sensitive to the assumptions on the probability distributions for the inputs and associated correlations, which are validated using different statistical tests (e.g. goodness-of-fit tests). As larger amounts of data become available, the use of simulation techniques that do not require any assumption on probability distributions are becoming more popular (e.g. bootstrapping). Wavestrapping statistical technique is one of those methods.

The present paper explores the application of wavestrapping as a statistical/simulation technique in a data analysis problem associated with train aerodynamics. In such a problem, there is usually a need to estimate the peak value of slipstream air speed. This peak/extreme value in train aerodynamics needs to be limited, so that, for instance, no passenger in a given platform is in danger when a train passes without stopping in that platform. Of course, local conditions of the platform (buildings and other constructions around) also play a significant role in this peak/extreme value. Nevertheless, the

¹Corresponding author. IDMEC, Instituto Superior Técnico, Universidade de Lisboa, Portugal, Av. Rovisco Pais, 049-001 Lisboa, Portugal.

²Rail Safety and Standards Board (RSSB), The Helicon, 1 South Place, London EC2M 2RB, UK.

³ University of Huddersfield, Institute of Railway Research, Queensgate, Huddersfield HD1 3DH, UK.

current practice, provided by the European standard EN 14067, requires that 20 valid measurements are taken and filtered using a 1-second moving average filter, so that a permissible characteristic air speed (97.5th percentile) can be calculated and compared with the limit values, by assuming that the peak/extreme values in a slipstream air speed measurement follows a normal distribution. Such procedure has been validated for a larger number of measurements in Baker et al. (2014a, 2014b), as part of a larger research work integrated in the AeroTRAIN research project. Our work explores the use of wavestrapping technique to estimate the permissible characteristic air speed, without the need to assume any probability distribution and compare it with the current practice of the EN 14067.

The outline of the present paper is the following: this first section introduced the main topic and the need for doing research in simulation methods for data analysis of train aerodynamic problems; then section 2 provides a brief background on train aerodynamics relevant to the topic. Section 3 discusses the wavestrapping statistical technique. Then, section 4 provides a simple descriptive analysis of the aerodynamic data used in this study and section 5 applies the wavestrapping technique to the aerodynamic data. Finally, section 6 provides some conclusions and steps for further research.

2. Brief background on train aerodynamics

Research on train aerodynamics has benefitted from several contributions. Different approaches emerge, for example using a computational fluid dynamics (CFD) approach, in which computational models aim to reproduce the flow around trains (Sima et al. 2015; Sesma et al. 2012), or more empirical approaches using reduced-scale and full-scale measurements (Paradot et al. 2015), which are also used to validate the CFD approaches. This brief background only provides a discussion on what is considered relevant for the main topic of the paper.

A review on flow around trains, namely on high speed trains, and on simulation of unsteady aerodynamic cross wind forces on trains can be found in Baker et al. (2010a and 2010b), which provide an overview of the physical phenomenon associated with train aerodynamics. It is also shown that if this dynamic phenomenon is of the order of 1s or more, then quasi-steady calculations can be assumed, in particular for high speed trains (Baker 2010b). This important result provides the argument to pre-filter the slipstream air speed measurements using a 1-second moving average filter before analysing the peaks.

The AeroTRAIN research project carried out full-scale measurements in different trains and sites, and produced extensive analysis of train slipstreams and wakes, using an ensemble average analysis (Baker et al. 2014a) and a gust analysis (Baker et al. 2014b). A statistical analysis of the maximum gusts (or peak values) was carried out in Baker et al. (2014b), enabling to assess a standard uncertainty associated with the gust parameter, and showing that: for most trains, the peak occurs in the near wake of the train; but for double-unit trains, it can occur around the gap between the units, and for the locomotive/coach combinations, it can even occur around the nose of the locomotive or at the discontinuity between train and locomotive. The authors conducted a statistical analysis on the 1-second gusts for the S-103 train as they had a large sample size, and decided to split their sample into three groups in order to understand the nature of these results: i) the peak occurs along the length of the train (0-200m); ii) the peak occurs in the near wake (200-600m) and iii) the peak occurs in the far wake. Then, they compared the frequency distributions for each group with a Normal distribution with mean and standard deviations calculated from the sample, concluding that it is shown that they can

be represented by normal distributions. Nevertheless, the authors do not report any result from a goodness-of-fit statistical test to validate their hypothesis and, apparently rely exclusively on a comparison of histograms. For the second group this comparison might be acceptable (though this group has the larger size $n=255$), for the other two groups (first and third), the assumption that the peaks are normally distributed seems to be questionable. Although this might seem a somewhat ‘quixotic’ question, as there is the practical need to simplify procedures when technical specifications for interoperability (TSI) are being proposed/revised, this normality assumption might simplify too much the calculation of the permissible characteristic air speed (97.5th percentile), by reducing it to the estimation of the mean and the standard deviation, and disregarding more information contained in the measurements/signals.

In fact, the EN 14067-4 states that ‘the (full-scale) tests shall consist of at least 20 independent and comparable test samples measured with reference wind speeds not exceeding 2 m/s. The slipstream airflow measurement shall start at least 4 s before the train nose passes and continue until at least 10 s after the train tail has passed. For a valid set of measurements, at least 50% of the measurements shall be taken within $\pm 5\%$ of the nominal test speed and 100% of the measurements shall be within $\pm 10\%$ of the nominal test speed.’ Therefore, the EN 14067 provides the following expression to estimate the permissible characteristic air speed:

$$U_{95\%} = U_{2\sigma} = \bar{U} + 2\sigma$$

In which: $U_{95\%}$ is the permissible characteristic air speed, \bar{U} is the mean value over all measured maxima U_i and σ is the standard deviation of all measured maxima U_i .

Note that the expression above relies on the assumption that the peaks follow a normal distribution (as it was concluded in Baker et al. 2014b), without any statistical assumption on the each of the observed values X_{it} for measurement i at time t , where $U_i = \max_t(X_{it})$, but instead assuming that the permissible characteristic air speed would be approximately the upper bound of a 95% confidence interval of the peaks, leading to $\bar{U} + 1.96\sigma$, where $1.96 = \phi_z^{-1}(0.975)$, i.e. the inverse of cumulative probability function (ϕ_z^{-1}) for probability 0.975. In other words, the permissible air speed corresponds to the 97.5th percentile of the peak distributions.

Relevant research work has also been published by Bell et al. (2014, 2015, 2016a, 2016b) on the behaviour of the wake flow and slipstream in high speed trains. For instance, Bell et al. (2014) showed that the probability distribution of flow speed was approximately a normal distribution (without using any statistical test to support this hypothesis), but as we move in the streamwise direction it changes from a normal distribution to a decreasing kurtosis and increasing positive skewness (reminding a Log-Normal distribution). Moreover, gust analysis in scaled equivalents has been conducted and other methods have been used, such as the Proper Orthogonal Decomposition (POD) to attempt to gain insight into the run-run variation in the near wake (Bell et al. 2015, 2016a, 2016b). In fact, the authors were able to show that the dominant feature of the time-average wake topology consists of a clearly identifiable counter-rotating streamwise vortex pair (Bell et al. 2016a). These works provide an important step forward in the understanding of the unsteady wake and its effect on slipstream, which might be then used for design optimization in the next generation of high speed trains (Bell et al. 2016b).

3. Wavestrapping statistical simulation method

Wavestrapping is a sampling technique in the wavelet domain. It joins two interesting concepts in statistical signal processing: Wavelet Analysis (WA) and Bootstrapping. This technique was first proposed by Percival et al. (2000). It has been applied to the analysis and synthesis of various time series data including atmospheric turbulence (Angelini et al. 2005), solar irradiance (Perpinan and Lorenzo 2011) and magnetic resonance imaging (Bullmore et al. 2003, 2004, Tang et al. 2008). It is mainly appropriate for time series that exhibit long memory properties, i.e. that show correlation over long periods.

According to Ogden (1997), WA is 'roughly speaking' a refinement of Fourier Analysis (FA). FA allows the frequency components of a signal to be distinguished, whereas WA can cope with signals whose frequency content change over time. Wavelets are 'little waves' that are finitely extended and WA is a way to decompose a given time series in an orthonormal basis of these wavelets.

Following Bullmore et al. (2004), a Wavelet family can be obtained by dilating and translating a compactly supported "mother" wavelet ψ with zero integral over time $\int \psi(t)dt = 0$:

$$\psi_{j,k}(t) = \frac{1}{\sqrt{2^j}} \psi\left(\frac{t - 2^j \cdot k}{2^j}\right)$$

And by dilating and translating a "father" wavelet or scaling function ϕ with unit integral over time $\int \phi(t)dt = 1$:

$$\phi_{j,k}(t) = \frac{1}{\sqrt{2^j}} \phi\left(\frac{t - 2^j \cdot k}{2^j}\right)$$

In which: $j = 1, 2, 3, \dots, J_0$ indexes the scale $S_j = 2^j$ ($S_1 = 2, S_2 = 4, \dots, S_{J_0} = 2^{J_0}$) to which the wavelet has been dilated, and $k = 1, 2, 3, \dots, K_j$ indexes the location in time (or space) to which it has been translated (in which: $K_j = N/2^j$ so that $K_1 = N/2^1, K_2 = N/2^2, \dots, K_{J_0} = N/2^{J_0}$).

Let $y = \{y_t\}$ be an initial signal (or the original data), where $t = 0, 1, \dots, N - 1$, with $N = 2^J$ representing its length for some integer J . Let us also pick an integer J_0 (so that $J_0 < J$) that would define the maximum scale of analysis ($S_{J_0} = 2^{J_0}$). At each scale, the initial data is decomposed into two orthogonal components: the detail coefficients $d_{j,k}$ and the approximation coefficients $a_{J_0,k}$, so that the original data y_t can be expressed as:

$$y_t = \sum_{k=1}^{2^{J-J_0}} a_{J_0,k} \cdot \phi_{J_0,k}(t) + \sum_{j=1}^{J_0} \sum_{k=1}^{2^{J-j}} d_{j,k} \cdot \psi_{j,k}(t)$$

These coefficients $d_{j,k}$ and $a_{J_0,k}$ can be computed as the inner products of the original data (y) with the corresponding scaled and dilated mother and father wavelets, i.e. $d_{j,k} = \langle y, \psi_{j,k} \rangle$ and $a_{j,k} = \langle y, \phi_{j,k} \rangle$, using an efficient pyramid algorithm proposed in Mallat (1989). The detail coefficients $d_{j,k}$ will contain information about variation in the data at a particular scale and the approximation coefficients $a_{J_0,k}$ represent the residual of the data after the information on this and all finer scales has been removed.

Therefore, the central idea of wavestrapping is to sample with replacement (i.e. bootstrap) for each level j the coefficients $d_j = \{d_{j,k}\}$ as they are uncorrelated, and leave the approximated/scaling

coefficients $a_{J_0} = \{a_{J_0,k}\}$ unchanged, and reconstruct the time series corresponding to each combination of the new sampled coefficients.

Bootstrapping is a technique that according to Robert and Casella (2004) was first proposed by Efron (1979). Bootstrapping is resampling with replacement from the observed data, and makes use of the resampled sequences to evaluate the properties of a given estimator through its empirical distribution, rather than assuming or using a theoretical distribution (e.g. normal distribution). In wavestrapping, we are not resampling from the original data in the time domain, but from a set of coefficients that correspond to the transformation of the original data in the wavelet domain.

It should be noted that wavestrapping resolution, i.e. the choice of J_0 , would adapt to ‘how correlated’ the data is. In fact, it is a little bit more complicated than that, as it makes use of Discrete Wavelet Packet Transforms (DWPT), instead of simple Discrete Wavelet Transforms. The DWPT generalize the Discrete Wavelet Transforms from the scale domain to the frequency domain, by computing Fourier Transforms of the “mother” and “father” wavelets. The wavestrapping method using the DWPT will initially search for the less correlated (‘the whitest’) transform of a DWPT, where the goal is to segment the normalized frequency interval $f \in [0, 1/2]$ into subintervals such that each segment of the Spectral Density Function (SDF) is as flat as possible within each subinterval. The term “white noise” (or whitest) is used to refer to uncorrelated random samples, which are Gaussian distributed, and thus independent and identically distributed. For further details, the reader is remitted to wavestrapping references (Percival et al. 2000, Tang et al. 2008), and the R package ‘wmtsa’ (Constantine and Percival 2014).

4. Aerodynamic data

This section provides an overview on the aerodynamic data, which was generously provided by AeroTRAIN project and it was extensively reported on Baker et al. (2014a, 2014b), and here is used to exemplify the wavestrapping technique. The data analysed contains 20 time-series measurements of the slipstream resultant air speed (in m/s). The measurement/sampling frequency is 600 per second, and most of the 20 measurements are in the interval (0, 30s), except two measurements (numbers 13 and 14) that are in the interval (0, 35s). The data was already pre-filtered using a 1-second moving average filter, though it contained 600 values per second it had been previously resampled at 10Hz. Note that in the following sections, the different time series (or signals) are referred to as measurements (or time-series measurements) as the term ‘measurements’ is more common in train aerodynamic literature.

The train tested is the S-103, which is a high-speed train constructed by Siemens and operated in Spain by RENFE. It runs as an eight-car set, with a driving unit at each end and has a top service speed of 310 km/h. It is 200.3 m long and travels at a nominal speed of 80 m/s (around 290 km/h or 180 mph), so it will take approximately 2.5 s to pass. Figure 1 provides an example of a measurement (number 1) of the slipstream resultant air speed at two different heights above the top of the rail ($H=0.20$ m and $H=1.20$ m). This graph shows a typical behaviour with two interesting peaks in each measurement: i) the first peak (P_1) is associated with the train nose passing and ii) the highest peak (P_{max}) is associated with the effect of the train tail.

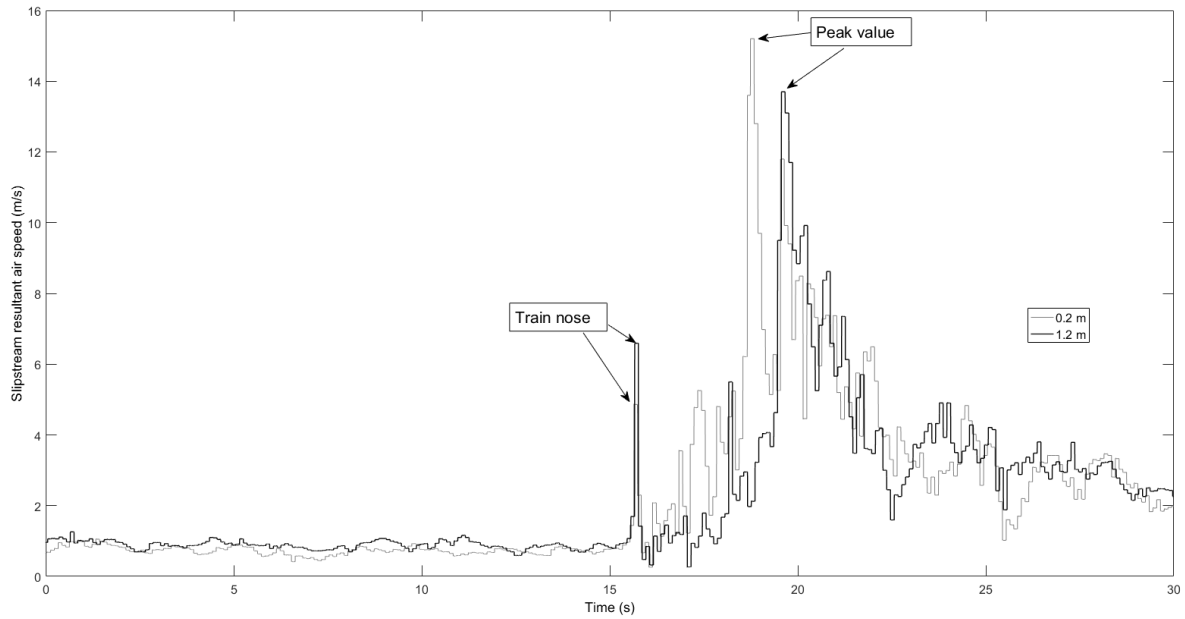
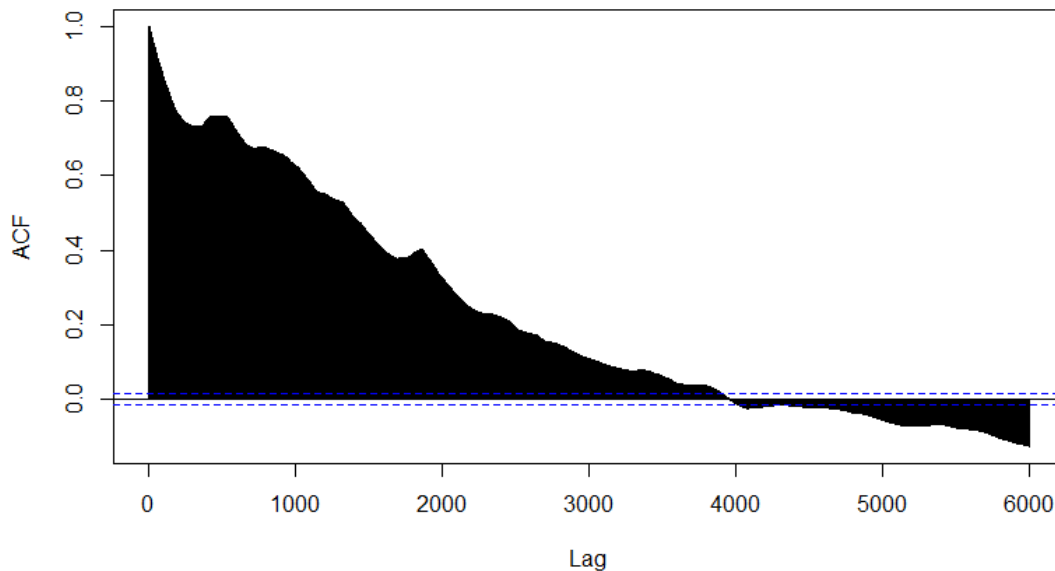
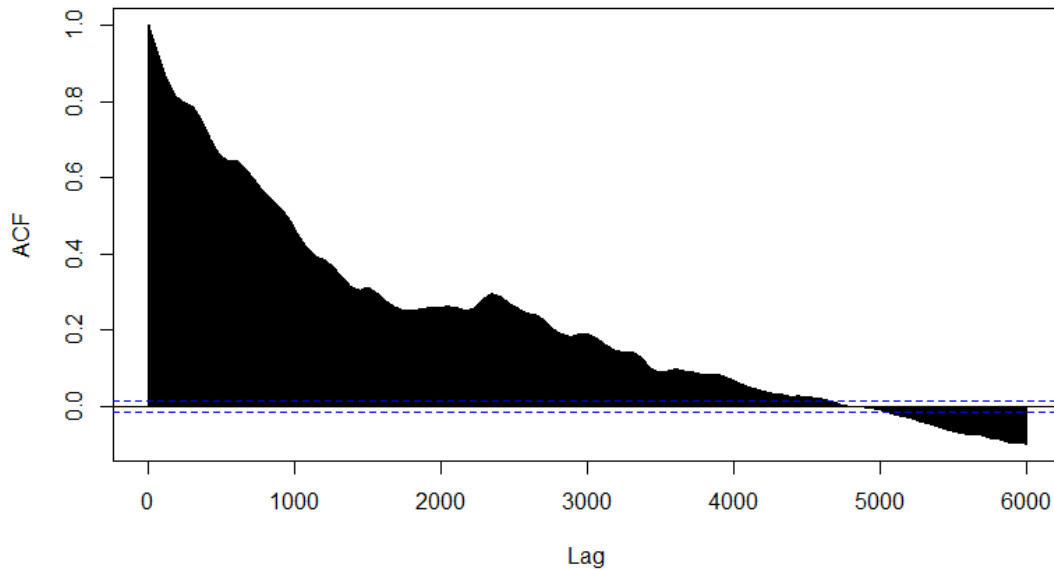


Figure 1 – Slipstream resultant air speed for a typical measurement (number 1) in the time interval (0, 30s) at two different heights $H=0.2\text{m}$ and $H=1.2\text{m}$.

The first question to be addressed is whether wavestrapping is an appropriate statistical technique to apply to this type of measured data. As it is highlighted in Percival et al. (2000), wavestrapping is particularly appropriate for sample signals that exhibit a long memory property, i.e. signals that exhibit an Autocorrelation Function (ACF) which is consistently statistical significant for different lags. Figure 2 shows the ACF for the maximum lag of 6000 (corresponding to a maximum time interval lag of 10s) for both signals at different heights $H=0.2\text{m}$ and $H=1.2\text{m}$.

The two peaks at height $H=0.2\text{m}$ are likely to be associated with vortices shed from the lowest part of the bogies regions, whereas at the higher height the train geometry is smoother. One of these vortices may not be recorded here, as the ballast shoulder could be influencing the $H=0.2\text{m}$ measurements.





Figures 2 – Autocorrelation Functions (ACF) for a typical measurement (number 1) at different heights: H=0.2 m (above) and H=1.2 m (below).

Figures 2 show that the measurement signals are correlated for different lags, and only after 4000 points (for H=0.2m top figure) or 5000 points (for H=1.2m bottom figure), do they exhibit an autocorrelation close to zero. This confirms that these signals exhibit the long memory property, for which wavestrapping is considered appropriate.

Figures 3 and 4 show the mean, the maxima and the minima of the slipstream resultant air speed that the 20 real measurements exhibit in the time interval (0, 30s), at heights H=0.2m and H=1.2m, respectively.

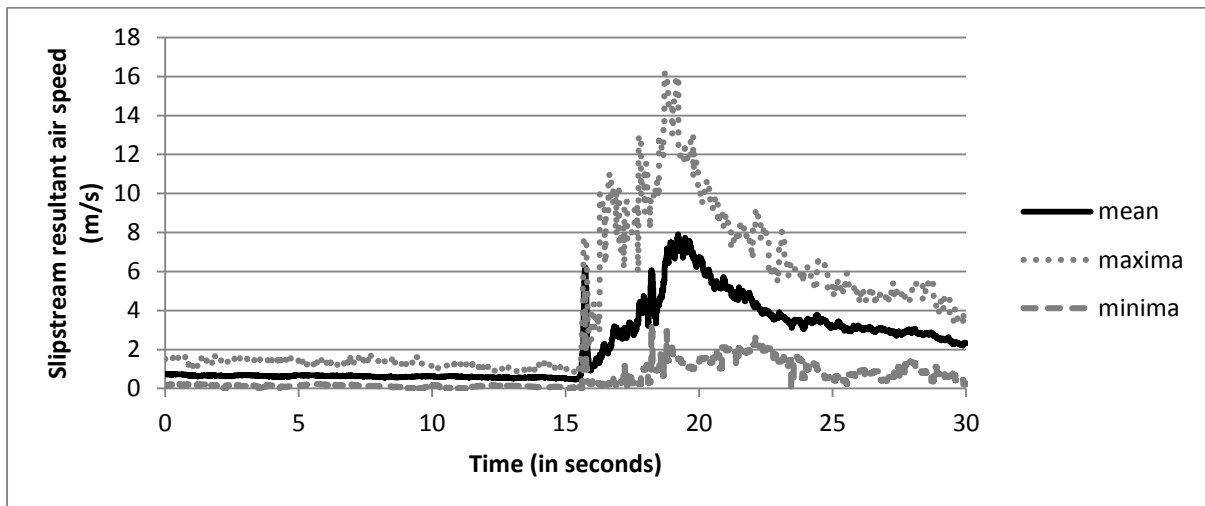


Figure 3 – Mean, maxima and minima of the slipstream resultant air speed for the 20 real measurements at height H=0.2 m in the time interval (0, 30s).

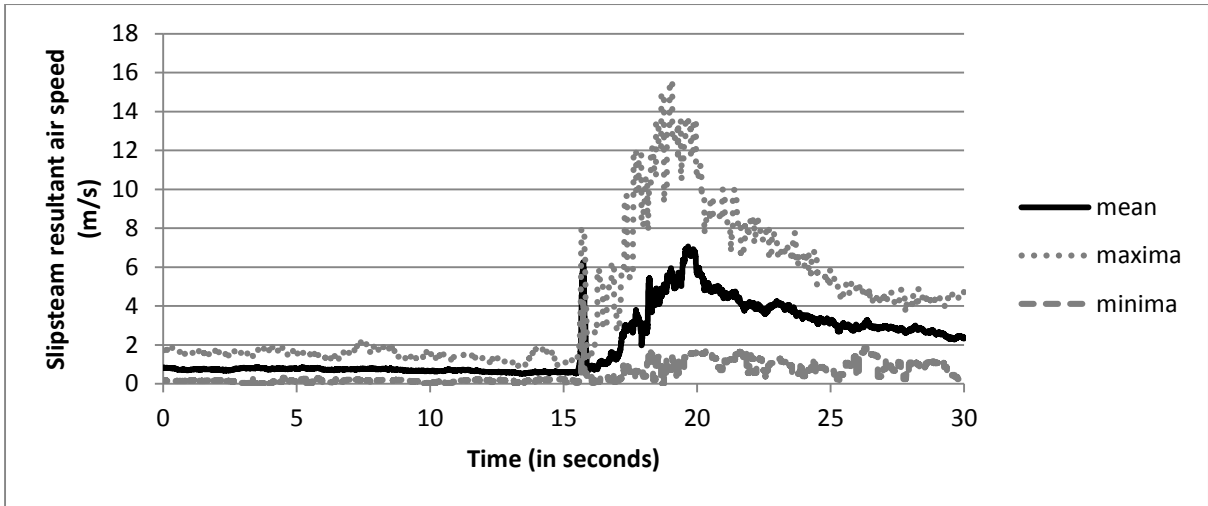


Figure 4 – Mean, maxima and minima of the slipstream resultant air speed for the 20 real measurements at height H=1.2 m in the time interval (0, 30s).

Note that the first 15s of the measured data, prior to the arrival of the train at the measuring location point, contains only a low and relatively constant airspeed. This part of the signal is not considered of interest and thus, the wavestrapping statistical technique is applied on the data measured after 15s.

5. Wavestrapping aerodynamic data

To test the validity of the wavestrapping method to estimate the permissible characteristic air speed ($U_{95\%}$), the following steps were proposed:

- 1) Use all the 20 measurements and compute the permissible characteristic air speed ($U_{95\%}$) according to EN 14067 (**current method**);
- 2) Use each measurement and sample 20, 50, 100, 1000, 5000, 10000, 100000 and 500000 signals using the wavestrapping technique (**wavestrapped/sampled series**);
- 3) Compare estimations for the permissible characteristic air speed ($U_{95\%}$) using the standard approach and the wavestrapping technique (**comparison between (1) and (2)**).

5.1. Current method based on standard EN 14067

The EN 14067 (2013) requires that the tests consist of at least 20 independent and comparable test samples measured with reference wind speeds not exceeding 2 m/s. Moreover, the slipstream airflow measurement shall start at least 4 s before the train nose passes and continue until at least 10 s after the train tail has passed. For a valid set of measurements, at least 50 % of the measurements shall be taken within $\pm 5\%$ of the nominal test speed and all of the measurements shall be within $\pm 10\%$ of the nominal test speed. All the measurements complied with these requirements.

Table 1 provides the values of slipstream resultant air speed for the main peak (P_{max} or U_i) for all measurements for the two different heights $H=0.2$ and $H=1.2$ m, as well as the calculated means, the standard deviations and the permissible characteristic air speeds for the two different heights ($U_{95\%}$).

P_{max}- Maximum peak due to train tail		
Measurement	H=0.2 m	H=1.2 m
	(m/s)	(m/s)
1	15.212	13.680
2	11.919	13.488
3	12.016	11.344
4	12.304	9.601
5	10.147	8.533
6	11.257	14.931
7	14.579	12.014
8	13.456	10.512
9	11.004	5.842
10	12.802	7.725
11	11.194	11.546
12	16.202	10.615
13	10.966	8.762
14	7.292	8.029
15	11.533	12.689
16	9.734	7.786
17	8.699	8.849
18	11.811	10.507
19	15.963	15.523
20	11.800	10.236
Mean (\bar{U})	11.995	10.611
St. Dev. (σ)	2.275	2.574
$U_{95\%} (\bar{U} + 2\sigma)$	16.545	15.760

Table 1 – Peak values due to train tail (P_{max} or U_i) for different heights above the top of the rail (H=0.2 m and H=1.2 m). Values are in m/s.

5.2. Using wavestrapping

The following results were obtained using the wavestrapping toolkit from R package statistical software 'wmtsa' (Constantine and Percival 2014), developed for Time Series Analysis.

As many applications using wavestrapping assume that the initial signal has length equal to $N = 2^J$ for some integer J . As the data was pre-filtered using a 1-second moving average filter, data was 'downsampled', with each measurement run having a maximum of 18000 or 21000 points, i.e. a time interval of 30 or 35 seconds, and after 'downsampling' (taking each 30th value), leads to a maximum of 600 or 700 points, corresponding to a maximum length of measurements of base two equal to $2^9 = 512$ observations. Note that $2^{10} = 1024$ observations, which is higher than 600 and 700, and thus $2^9 = 512$ is the maximum number of observations (as a power of two: $N = 2^J$) that can be contained in each measurement/run. Nevertheless, as it was shown in Figure 1, each measurement will have two quite different characteristic regions: the first region before the train nose has passed (almost constant), and a second region after the train nose has passed (with a lot of variability). Therefore, it was decided to exclude the initial values from the measurement 0.5s before the train nose passed, leading to an effective interval of analysis from (15.1s, 27.9s), which represent a total of 256 observations at each 0.05s, and thus $J = 8$.

Therefore, the following analysis only used the subinterval (15.1s, 27.9s) 'downsampled' at each 0.05s with a total of 256 observations. A useful way to analyse the contribution of each frequency subinterval to the original signal is plotted in Figure 5. It shows the reconstructed signal if only that particular vector of Discrete Wavelet Transforms (for that particular frequency subinterval) is used to reconstruct the signal. For level $j = 0$, the original signal is represented in blue ($W_{0,0} \equiv X$), whereas for the lower levels ($j = 1, 2, 3, 4$), we can see that several squares contribute almost nothing to the original signal, namely those on the right side (i.e. for higher frequencies). For instance, note that all reconstructed signals are supposed to represent the same original extension as the original signal at level 0, but they are shrunk as it is symbolised by green arrows, so that all reconstructed signals at a given level can fit the width of the figure.

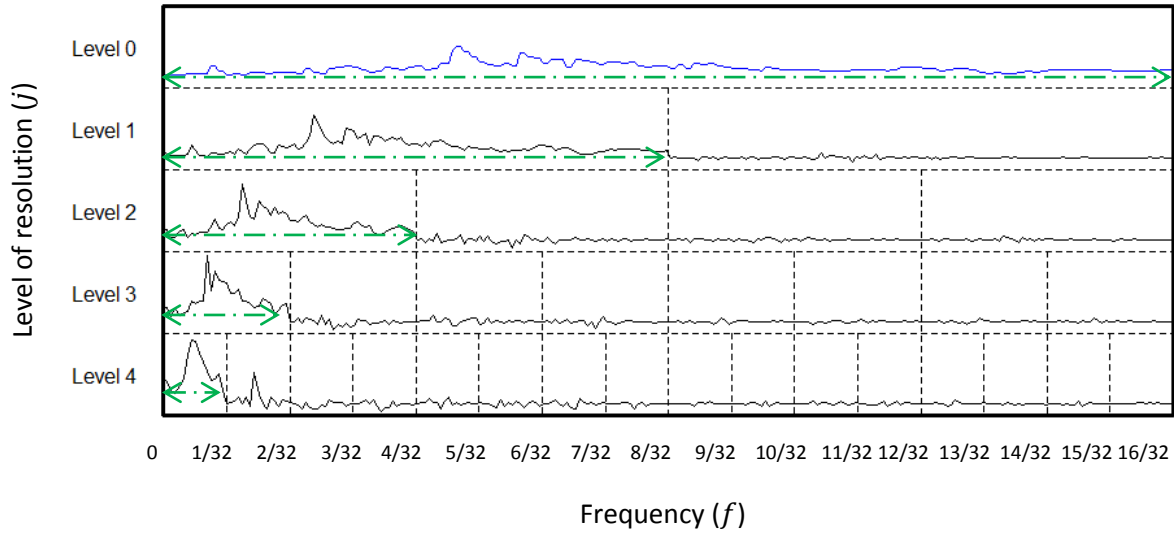


Figure 5 – Discrete Wavelet Packet Transforms of a typical measurement (number 1) at height H=0.2, for the interval (15.1s, 27.9s) 'downsampled' at 0.05s, using the Daubechies wavelet 's8' as the wavelet family, for the level $J_0 = 4$.

Another useful plot is in Figure 6, which shows the detail coefficients $d_j = \{d_{j,k}\}$ for each level j and the approximated/scaling coefficients $a_{J_0} = \{a_{J_0,k}\}$. Figure 6 exhibits these coefficients for the DWT using as the wavelet family the Daubechies wavelet 's8' and for level $J_0 = 4$.

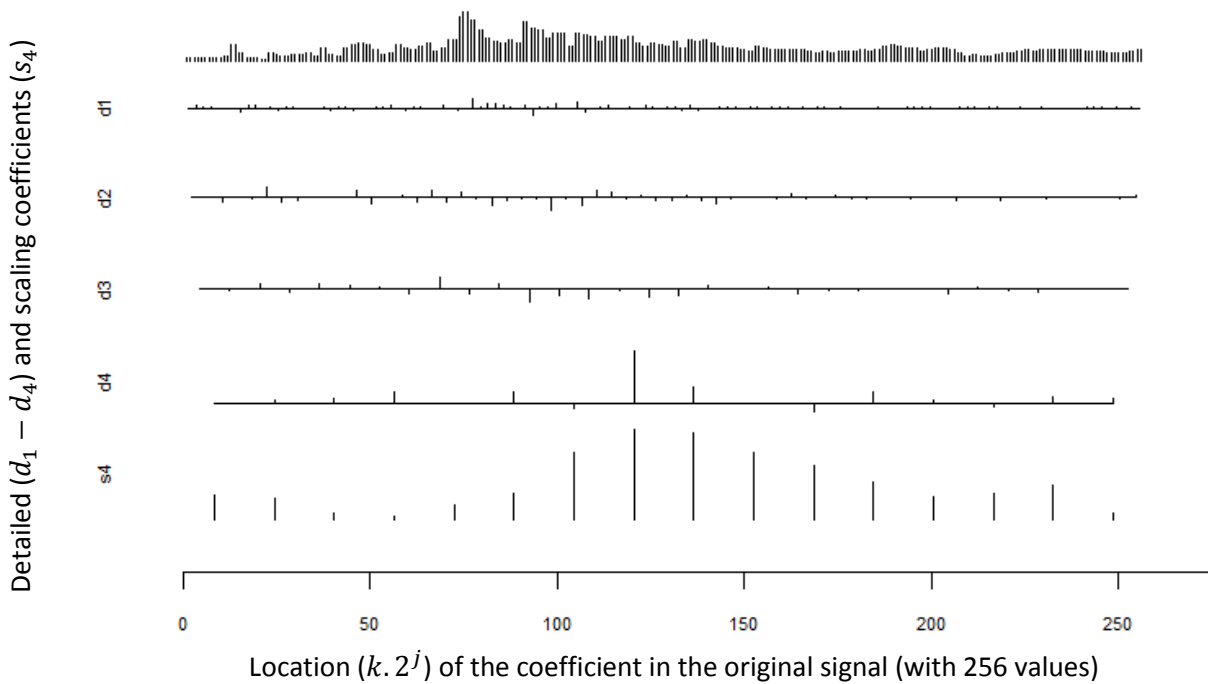


Figure 6 – Detailed and scaling coefficients using the Daubechies wavelet family 's8' for level $J_0 = 4$ for the 256 values of measurement number 1 (at height H=0.2m) for the time interval (15.1s, 27.9s) 'downsampled' at 0.05s.

From Figure 6, we can see that the main contribution of the signal is given by the scaling coefficients (a_4 or s_4 as it is given in Figure 6). In fact, Table 2 quantifies some statistics of the detailed and scaling coefficients and the percentage of contribution at each level, showing that the scaling coefficients

provide the greater part of the energy of the signal. Note that as the wavestrapping method is sampling only the detailed coefficients at each level j , leaving the scaling coefficient unchanged, if J_0 would be chosen equal to 4, then 90.9% of the signal energy is actually not resampled.

	Min (m/s)	Max (m/s)	Mean (m/s)	SD (m/s)	VAR (m ² /s ²)	%Energy
d_1	-2.342	2.644	0.038	0.530	0.281	0.608
d_2	-3.883	3.137	-0.116	1.150	1.323	1.429
d_3	-4.218	3.535	-0.329	1.525	2.327	1.283
d_4	-2.767	16.214	2.021	4.302	18.504	5.821
s_4	5.600	32.897	16.197	8.777	77.041	90.859

Table 2 – Main statistics and contribution of the detailed and scaling coefficients in Figure 6.

A natural question is which level J_0 (maximum level of resolution) should be chosen to decompose the initial signal. The wavestrapping algorithm chooses the ‘whitest’ decomposition of DWPT crystals $W_{j,n}$ with a maximum J_0 equal to $J - 2$. Table 3 present the ‘whitest’ decompositions for all 20 measurements at height $H=0.2m$, using the Portmanteau test ‘type I’ with a significance level of 5%. Note that the term ‘crystals’ is here used following Constantine and Percival (2014), which refers to the vectors of the wavelet transform coefficients. For further details, the reader is remitted to the R package ‘wmtsa’ and supporting documentation (Constantine and Percival 2014).

A simple analysis of the predominance of some DWPT crystals over others in the whitest decompositions for each measurement signal at height $H=0.2m$ would provide an idea of whether the approximation of white noise in that particular subinterval of frequency is consistent. The DWPT crystal appearing more times in the whitest decompositions are: $W_{5,1}$ (20 out of 20), $W_{4,1}$ (19 out of 20), $W_{6,0}$ (18 out of 20), $W_{6,1}$ (18 out of 20) and $W_{3,1}$ (16 out of 20). As almost all ‘whitest’ decompositions reach the maximum level $J_0 = J - 2 = 8 - 2 = 6$, with crystals $W_{6,0}$ and $W_{6,1}$, this may provide an indication that in the frequency subintervals $(0, 1/128)$ and $(1/128, 2/128)$, the reconstructed signals reject normality (white noise assumptions) and we would need more points (than the 256 values) in the original signal to keep subdividing these frequency subintervals till white noise assumption become valid. Further research on the exploration of the predominance of some DWPT crystals over others is needed. This is out of the scope of the present paper.

Measurement ($H=0.2m$) ‘Whitest’ decomposition

- 1 $W_{2,1}, W_{2,2}, W_{2,3}, W_{3,1}, W_{4,1}, W_{5,1}, W_{6,0}, W_{6,1}$
- 2 $W_{2,1}, W_{2,3}, W_{3,1}, W_{3,4}, W_{4,1}, W_{4,10}, W_{4,11}, W_{5,1}, W_{6,0}, W_{6,1}$
- 3 $W_{2,1}, W_{2,2}, W_{3,7}, W_{4,1}, W_{4,2}, W_{4,3}, W_{4,13}, W_{5,0}, W_{5,1}, W_{5,25}, W_{6,48}, W_{6,49}$
- 4 $W_{2,1}, W_{2,3}, W_{3,2}, W_{3,3}, W_{4,2}, W_{4,3}, W_{5,1}, W_{5,2}, W_{5,3}, W_{6,0}, W_{6,1}$

5	$W_{2,1}, W_{2,2}, W_{3,1}, W_{3,6}, W_{4,1}, W_{4,14}, W_{4,15}, W_{5,1}, W_{6,0}, W_{6,1}$
6	$W_{3,1}, W_{3,2}, W_{3,3}, W_{3,4}, W_{3,7}, W_{4,1}, W_{4,10}, W_{4,11}, W_{4,12}, W_{4,13}, W_{5,1}, W_{6,0}, W_{6,1}$
7	$W_{3,1}, W_{3,2}, W_{3,3}, W_{3,5}, W_{3,6}, W_{3,7}, W_{4,1}, W_{4,8}, W_{4,9}, W_{5,1}, W_{6,0}, W_{6,1}$
8	$W_{2,1}, W_{2,2}, W_{3,1}, W_{3,6}, W_{3,7}, W_{4,1}, W_{5,1}, W_{6,0}, W_{6,1}$
9	$W_{2,1}, W_{2,3}, W_{3,1}, W_{3,4}, W_{3,5}, W_{4,1}, W_{5,1}, W_{6,0}, W_{6,1}$
10	$W_{2,1}, W_{2,2}, W_{2,3}, W_{3,1}, W_{4,1}, W_{5,1}, W_{6,0}, W_{6,1}$
11	$W_{2,2}, W_{3,2}, W_{3,6}, W_{4,1}, W_{4,2}, W_{4,3}, W_{4,6}, W_{4,7}, W_{4,14}, W_{4,15}, W_{5,01}, W_{6,0}, W_{6,1}$
12	$W_{3,1}, W_{3,3}, W_{3,4}, W_{3,6}, W_{3,7}, W_{4,1}, W_{4,4}, W_{4,5}, W_{4,10}, W_{4,11}, W_{5,0}, W_{5,1}$
13	$W_{1,1}, W_{3,2}, W_{3,3}, W_{4,1}, W_{4,3}, W_{5,1}, W_{5,4}, W_{5,5}, W_{6,0}, W_{6,1}$
14	$W_{3,1}, W_{3,2}, W_{3,5}, W_{3,6}, W_{3,7}, W_{4,1}, W_{4,6}, W_{4,7}, W_{4,8}, W_{4,9}, W_{5,01}, W_{6,0}, W_{6,1}$
15	$W_{2,2}, W_{3,1}, W_{3,3}, W_{3,6}, W_{4,1}, W_{4,4}, W_{4,14}, W_{4,15}, W_{5,1}, W_{5,10}, W_{5,11}, W_{6,0}, W_{6,1}$
16	$W_{2,2}, W_{3,1}, W_{3,2}, W_{3,3}, W_{3,6}, W_{3,7}, W_{4,1}, W_{5,1}, W_{6,0}, W_{6,1}$
17	$W_{2,2}, W_{2,3}, W_{3,1}, W_{4,1}, W_{4,4}, W_{4,5}, W_{4,6}, W_{4,7}, W_{5,1}, W_{6,0}, W_{6,1}$
18	$W_{2,1}, W_{2,2}, W_{2,3}, W_{3,1}, W_{4,1}, W_{5,1}, W_{6,0}, W_{6,1}$
19	$W_{1,1}, W_{2,1}, W_{3,1}, W_{4,1}, W_{5,1}, W_{6,0}, W_{6,1}$
20	$W_{2,1}, W_{2,2}, W_{3,1}, W_{3,7}, W_{4,1}, W_{4,12}, W_{4,13}, W_{5,1}, W_{6,0}, W_{6,1}$

Table 3 – ‘Whitest decompositions of the measurement signals for different runs at height H=0.2m.

The wavestrapping algorithm will then choose the ‘whitest’ decomposition for a particular initial measurement, and then sample with replacement different detailed coefficients at each level. Figure 7 provides three examples of ‘wavestrapped’ time series \mathbf{b}_1 , \mathbf{b}_2 and \mathbf{b}_3 using measurement number 1 at height H=0.2 m as the initial measurement. Only data between 15 – 30s ‘downsampled’ to 256 values is considered.

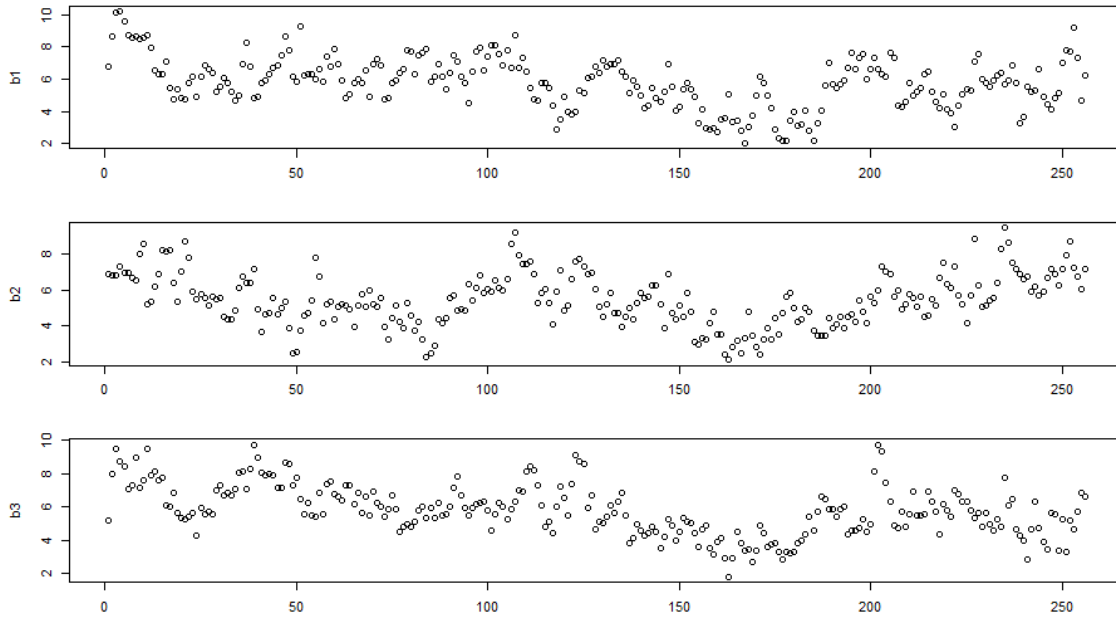


Figure 7 – Three waverstrapped time series b_1 , b_2 and b_3 with respective peaks equal to $P_1=10.218$, $P_2=9.427$ and $P_3=9.740$, with 256 values associated with the time interval (15.1s, 27.9s).

One important remark is that the fundamental characteristic of the slipstream has been lost, i.e. two distinct peaks due to the train nose and tail, which indicates that further work needs to be done to address the issue of simulating representative measurements that preserve this fundamental characteristic while waverstrapping non-fundamental fluctuations. Nevertheless, for the purpose of estimating the air permissible speed or the peak value, simulated series do not need to mimic those two peaks as long as they reproduce the desired wavelet shapes from typical signals.

As we waverstrapped these three series/signals b_1 , b_2 and b_3 , we can then compute the peak values for each time series, leading to $P_1=10.218$, $P_2=9.427$ and $P_3=9.740$. In the same way that these three signals were waverstrapped, we can sample as many waverstrapped time series as we want, or at least as long as the PC memory allows. Table 4 shows the main statistics of the peak values (mean, standard deviation, and some percentiles computed according to the standard $U_{95\%}$ and 95% and 97.5% percentiles from the sampled empirical distributions). Note that all main statistics seem to converge as N increases.

	N							
H=0.2m	20	50	100	1000	5000	10000	100000	500000
Mean (\bar{U})	10.588	10.766	10.747	10.646	10.782	10.759	10.742	10.743
St. Dev. (σ)	1.768	1.958	1.687	1.958	1.927	1.919	1.919	1.921
$U_{95\%} (\bar{U} + 2\sigma)$	14.125	14.681	14.122	14.561	14.635	14.597	14.581	14.586
$P_{95\%}$	13.248	13.783	13.826	13.983	14.169	14.154	14.123	14.129
$P_{97.5\%}$	13.382	14.182	14.416	14.576	14.632	14.729	14.657	14.643
	N							
H=1.2m	20	50	100	1000	5000	10000	100000	500000
Mean (\bar{U})	10.080	9.765	9.916	9.887	9.872	9.876	9.849	9.848
St. Dev. (σ)	1.817	1.542	1.511	1.461	1.473	1.470	1.467	1.464
$U_{95\%} (\bar{U} + 2\sigma)$	13.714	12.849	12.938	12.809	12.817	12.817	12.783	12.777
$P_{95\%}$	12.500	11.904	12.254	12.370	12.305	12.328	12.258	12.260
$P_{97.5\%}$	13.181	13.092	12.847	12.950	12.793	12.738	12.736	12.730

Table 4 – Asymptotic behaviour of some statistics as the number of wavestrapped time series N increases for two cases: height H=0.2m and height H=1.2m, using measurement number 1.

5.3. Comparing current method and wavestrapping

If we do the same procedure for all the other 20 measurements, i.e. use wavestrapping to calculate the 97.5th percentile from the empirical distribution (using for instance N=1000 wavestrapped time series for each measurement to estimate the peak distribution), we can then compare the estimations for the permissible characteristic air speed ($U_{95\%}$) calculated using the EN 14067 current method (sub section 5.1) with those calculated using the wavestrapping method (subsection 5.2).

Table 5 shows the main statistics and the permissible air speed and 95th and 97.5th percentiles for using N=1000 wavestrapped time series, assuming as initial signal (measurement number 1 to number 20). The permissible air characteristic speed ($U_{95\%}$) at height H=0.2m assumes the highest value for measurement number 19 (15.395), and at height H=1.2m assumes the highest value for measurement number 6 (15.108). Note that these two values are still lower than the values computed using the current method, respectively 16.545 and 15.760 (from Table 1). Another important remark is that the estimation of the permissible air characteristic speed is highly dependent on the measurement number used, reaching a lowest value of 7.844 for measurement number 14 at H=0.2m and the value of 7.273 for measurement 9 at height H=1.2m. Therefore, though the wavestrapped time series based on a single measurement can represent the variability within that single measurement, the wavestrapped time series from a single measurement cannot mimic the variability in a set of 20 real measurements.

Initial signal (H=0.2 m)																				
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Mean (\bar{U})	10.781	10.759	9.779	9.682	8.701	9.571	11.388	11.771	9.721	7.807	9.780	8.421	8.842	6.160	10.264	7.982	7.749	8.701	11.895	9.913
St. Dev. (σ)	1.914	1.214	1.656	1.400	0.851	1.294	1.683	1.199	1.347	1.057	1.010	1.802	1.133	0.842	1.140	0.849	0.936	1.516	1.750	1.411
$U_{95\%} (\bar{U} + 2\sigma)$	14.610	13.187	13.092	12.481	10.403	12.159	14.754	14.170	12.416	9.921	11.800	12.024	11.107	7.844	12.544	9.680	9.621	11.733	15.395	12.734
$P_{95\%}$	14.191	12.840	12.539	11.874	10.236	11.607	13.944	13.865	11.880	9.680	11.272	11.155	10.840	7.633	12.242	9.388	9.250	11.299	14.664	12.118
$P_{97.5\%}$	14.680	13.231	13.015	12.522	10.675	12.109	14.438	14.465	12.252	9.992	11.586	11.830	11.182	7.854	12.586	9.711	9.606	11.744	15.658	12.507
Initial signal (H=1.2 m)																				
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Mean (\bar{U})	9.896	11.723	8.338	8.542	7.923	11.389	10.874	10.574	5.798	6.257	9.145	6.425	8.456	5.950	10.891	6.632	7.785	7.808	11.197	8.528
St. Dev. (σ)	1.496	1.400	1.481	0.725	0.901	1.860	1.495	1.260	0.738	0.732	1.079	1.192	1.064	0.765	1.184	0.794	0.980	1.099	1.135	1.208
$U_{95\%} (\bar{U} + 2\sigma)$	12.888	14.523	11.299	9.992	9.725	15.108	13.864	13.095	7.273	7.720	11.304	8.809	10.585	7.480	13.260	8.219	9.746	10.005	13.467	10.943
$P_{95\%}$	12.431	13.562	10.870	9.771	9.314	14.535	13.116	12.688	7.049	7.474	10.923	8.537	10.210	7.306	12.767	7.982	9.451	9.672	13.163	10.514
$P_{97.5\%}$	12.909	13.892	11.197	10.029	9.576	15.140	13.518	13.116	7.346	7.766	11.438	9.040	10.648	7.571	13.220	8.281	9.949	10.065	13.636	11.001

Table 5 – Main statistics of N=1000 wavestrapped time series, using as original data different measurements (numbers 1 up to 20).

0 Nevertheless, one could then take the highest value for the 97.5th percentile (calculated for the
1 wavestrapped series using each single measurement) across all initial measurements and use that
2 value as an estimate of the peak value of slipstream air speed. In that way, one would get 15.658 for
3 $H=0.2$ m (measurement number 19) and 15.140 for $H=1.2$ m (measurement number 6). Note that
4 those values compare with 16.545 and 15.760 respectively (see Table 1), which might indicate that
5 empirical distributions have right-hand tails less heavy than the assumed normal distributions.

6 **6. Conclusions and further research**

7 This paper explored the application of wavestrapping to estimate an extreme value in train
8 aerodynamics. It compared two methods to estimate the permissible air characteristic speed: i) using
9 the current method from EN 14067 and ii) using a new statistical sampling technique called
10 wavestrapping. Wavestrapping as a simulation technique is suited for signals that exhibit long memory
11 properties, i.e. signals that show autocorrelation over long periods.

12 This study showed that the wavestrapping statistical technique can simulate different signals based
13 on a single measurement, and in this way, represent the variability within that single measurement.
14 Nevertheless, a single measurement cannot reproduce the variability of 20 real measurements.
15 Wavestrapping provided an alternative way to estimate a peak value that takes into account the
16 wavelet shapes of the aerodynamic signal and does not rely in normal distribution assumptions, as
17 the current EN 14067 method.

18 Further research might include combining wavestrapping with other filtering techniques to preserve
19 some fundamental characteristics of the signal and let the observed differences between the filtered
20 signal and the observed signal be wavestrapped. Moreover, possible paths for further research might
21 include an analysis of the performance of 'wavestrapping' versus more well-known sampling
22 techniques, used for instance in ARIMA (Autoregressive Incremental Moving Average models)
23 modelling approaches that would have a more straightforward sampling. A further refinement would
24 consist of modelling both measurements $H=0.2$ m and $H=1.2$ m at the same time, controlling for cross
25 correlation between them, even within the wavestrapping framework.

26 **Acknowledgments**

27 The authors wish to thank Rail Safety and Standard Board (RSSB) for the support and Professor Chris
28 Baker for generously sharing a sample of the aerodynamic data from AeroTRAIN EU research project
29 with us.

30 **References**

- 31 Angelini, C., Cava, D., Katul, G., & Vidakovic, B. (2005). Resampling hierarchical processes in the
32 wavelet domain: A case study using atmospheric turbulence. *Physica D: Nonlinear Phenomena*, 207(1),
33 24-40.
- 34 Baker, C. J. (2010). The simulation of unsteady aerodynamic cross wind forces on trains. *Journal of*
35 *Wind Engineering and Industrial Aerodynamics*, 98(2), 88-99.

36 Baker, C. (2010). The flow around high speed trains. *Journal of Wind Engineering and Industrial*
37 *Aerodynamics*, 98(6), 277-298.

38 Baker, C. J., Quinn, A., Sima, M., Hoefener, L., & Licciardello, R. (2014a). Full-scale measurement and
39 analysis of train slipstreams and wakes. Part 1: Ensemble averages. *Proceedings of the Institution of*
40 *Mechanical Engineers, Part F: Journal of Rail and Rapid Transit*, 228(5), 451-467.

41 Baker, C. J., Quinn, A., Sima, M., Hoefener, L., & Licciardello, R. (2014b). Full-scale measurement and
42 analysis of train slipstreams and wakes. Part 2 Gust analysis. *Proceedings of the Institution of*
43 *Mechanical Engineers, Part F: Journal of Rail and Rapid Transit*, 228(5), 468-480.

44 Bell, J. R., Burton, D., Thompson, M., Herbst, A., & Sheridan, J. (2014). Wind tunnel analysis of the
45 slipstream and wake of a high-speed train. *Journal of Wind Engineering and Industrial*
46 *Aerodynamics*, 134, 122-138.

47 Bell, J. R., Burton, D., Thompson, M. C., Herbst, A. H., & Sheridan, J. (2015). Moving model analysis of
48 the slipstream and wake of a high-speed train. *Journal of Wind Engineering and Industrial*
49 *Aerodynamics*, 136, 127-137.

50 Bell, J. R., Burton, D., Thompson, M. C., Herbst, A. H., & Sheridan, J. (2016a). Flow topology and
51 unsteady features of the wake of a generic high-speed train. *Journal of Fluids and Structures*, 61, 168-
52 183.

53 Bell, J. R., Burton, D., Thompson, M. C., Herbst, A. H., & Sheridan, J. (2016b). Dynamics of trailing
54 vortices in the wake of a generic high-speed train. *Journal of Fluids and Structures*, 65, 238-256.

55 Bullmore, E., Fadili, J., Breakspear, M., Salvador, R., Suckling, J., & Brammer, M. (2003). Wavelets and
56 statistical analysis of functional magnetic resonance images of the human brain. *Statistical methods*
57 *in medical research*, 12(5), 375-399.

58 Bullmore, E., Fadili, J., Maxim, V., Şendur, L., Whitcher, B., Suckling, J., ... & Breakspear, M. (2004).
59 Wavelets and functional magnetic resonance imaging of the human brain. *Neuroimage*, 23, S234-S249.

60 Chan, K. S., & Ripley, B. (2012). *TSA: Time series analysis*. R package version, 1.01.

61 Constantine, W., & Percival, D. (2014). *WMTSA: Wavelet Methods for Time Series Analysis*, R package
62 version 2.0-0.

63 Efron, B. (1979). Bootstrap methods: another look at the jackknife. *The annals of Statistics*, 1-26.

64 EN14067-4 (2013) *Railway applications - Aerodynamics - Part 4: Requirements and test procedures for*
65 *aerodynamics on open track*.

66 Mallat, S. G. (1989). A theory for multiresolution signal decomposition: the wavelet representation.
67 *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 11(7), 674-693.

68 Ogden, T. (1997). *Essential wavelets for statistical applications and data analysis*. Springer.

- 69 Paradot, N., Grégoire, R., Stiepel, M., Blanco, A., Sima, M., Deeg, P., ... & Zanetti, G. (2015). Crosswind
70 sensitivity assessment of a representative Europe-wide range of conventional vehicles. *Proceedings*
71 *of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit*, 229(6), 594-624.
- 72 Percival, D. B., Sardy, S., & Davison, A. C. (2000). Wavestrapping time series: Adaptive wavelet-based
73 bootstrapping. *Nonlinear and Nonstationary Signal Processing*, 442-471.
- 74 Perpiñán, O., & Lorenzo, E. (2011). Analysis and synthesis of the variability of irradiance and PV power
75 time series with the wavelet transform. *Solar Energy*, 85(1), 188-197.
- 76 Robert, C. P., & Casella, G. (2004). *Monte Carlo statistical methods*. Springer.
- 77 Sima, M., Eichinger, S., Blanco, A., & Ali, I. (2015). Computational fluid dynamics simulation of rail
78 vehicles in crosswind: Application in norms and standards. *Proceedings of the Institution of*
79 *Mechanical Engineers, Part F: Journal of Rail and Rapid Transit*, 229(6), 635-643.
- 80 Tang, L., Woodward, W. A., & Schucany, W. R. (2008). Undercoverage of Wavelet-Based Resampling
81 Confidence Intervals. *Communications in Statistics—Simulation and Computation*[®], 37(7), 1307-1315.
- 82