

Appendix 2: Discrete integral approximation

2.1 Total volume

Using the Taylor expansion, and neglecting terms of order higher than one, we obtain the following approximation to V_t^i :

where we defined

$$\mu(C_i) \equiv \int_{C_i} dp,$$

i.e., the Lebesgue measure of the i th Voronoi cell, and

$$I_i \equiv \int_{C_i} \Delta p dp.$$

$I_i / \mu(C_i)$ is the mass centroid of the i th cell with the origin placed at point p_i . If the observed points are the centroids of the cells, these quantities are identically zero. We will denote the error of this approximation by

$$\tilde{\epsilon}_V = \tilde{V}_t^i - V_t^i = - \sum_{i=1}^N \int_{C_i} (\Delta p^T H_S(s_i) \Delta p + h.o.t) dp.$$

2.2 Section volume

This problem is equivalent to the previous one, if we restrict the domain of analysis to one-dimensional manifolds (in this case, lines).

The quantities of interest are the area of the surface along certain reference lines L_i which are given by:

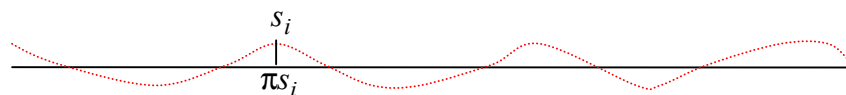
$$A_i = \int_{L_i} S(\ell) d\ell$$

where the set L_i is defined by

$$L_i = \left\{ \ell \in L_i : S(\ell) > z^{ref} \right\}.$$

where, for simplicity, we omit the dependency on time t .

Let $\{s_i\}_{i=1,\dots,N} \subset S_i$ be the observed points, and $\{\pi s_i\}_{i=1,\dots,N}$ their orthogonal projections on the line L_i



The integral can be approximated in terms of the values measured at positions $\{\pi s_i\}_{i=1,\dots,N}$ by using the Taylor expansion of the surface around each measured point:

$$S(\pi s_i + \ell) = S(\pi s_i) + \frac{\partial S(\pi s_i + \ell u)}{\partial \ell} \Big|_{\pi s_i} \ell + \frac{1}{2} \frac{\partial^2 S(\pi s_i + \ell u)}{\partial \ell^2} \Big|_{\pi s_i} \ell^2 + h.o.t.$$

where u is the unit vector along the direction of line L_i .

Let $\{\pi C_i\}_{i=1,\dots,N}$ be the set of Voronoï cells determined by the set of points $\{\pi s_i\}_{i=1,\dots,N}$. Using the Taylor expansion, and neglecting terms of order higher than one, we obtain the following approximation to A_i :

$$\begin{aligned} \tilde{A}_i &\equiv \sum_{j=1}^N \int_{\pi C_j} \left(S(\pi s_j) + \frac{\partial S(\pi s_j + \ell u)}{\partial \ell} \Big|_{\pi s_j} \ell \right) d\ell \\ &= \sum_{j=1}^N \left(S(\pi s_j) \mu(\pi C_j) + \frac{\partial S(\pi s_j + \ell u)}{\partial \ell} \Big|_{\pi s_j} I_j \right) \end{aligned}$$

where we defined

$$\mu(\pi C_i) \equiv \int_{\pi C_i} d\ell,$$

i.e., the Lebesgue measure of the i th cell, and

$$I_i \equiv \int_{\pi C_i} \ell d\ell.$$

$I_i / \mu(\pi C_i)$ is the mass centroid of the i th cell with the origin placed at point πs_i . If the observed points are the centroides of the cells, these quantities are identically zero.

We will denote the error of this approximation by

$$\tilde{\varepsilon}_A = \tilde{A}_i - A_i = - \sum_{j=1}^N \int_{\pi C_j} \left(\frac{1}{2} \frac{\partial^2 S}{\partial \ell^2} \Big|_{\pi s_i} \ell^2 + h.o.t. \right) d\ell.$$

Appendix 3: Sand bank model using wavelets

In this appendix we presented a study conducted using a set of profiles for three DECA lines provided by MUMM, on the possibility of identifying a reduced-dimensionality basis for the shape of the Kwintebank.

A particularly attractive basis for representing the profiles is given by wavelet theory, which have been successfully used to analyse non-stationary signals, whose frequency contents vary along time. It expresses a signal as a linear combination of basis functions which are well localised both in time and frequency. The most standard wavelet representations consider the expansion of a signal on a tree of basis functions, with increasing support and decreasing bandwidth, see [Daubechies] for more details.

We computed this representation for the Kwintebank data provided by MUMM, consisting of surveys of 3 distinct DECA lines (rG19, rG21 and rH01), comprising 25 surveys for each line, in a total of 75 profiles. We considered for this study the aligned and reconstructed profiles built as described in Deliverable D1.A (we have to give a name to this document...!).

3.1 Reduced dimensionality representation of the shape bank.

We use the wavelet basis determined by the mother wavelet “Symlets”, which are compactly supported wavelets with least asymmetry and highest number of vanishing moments for a given support width [Daubechies, Ten lectures of wavelets, CBMS, SIAM, 61, 1194, 194-202].

Let $\{S_i^k(\ell_n) \mid \ell_n \in L_k\}_{i=1}^{N_s}$ be the set of profiles for DECA line L_k , ordered in increasing time. In our case $N_s = 25$.

We computed the representation of the profiles in a complete wavelet tree of level 9, which allows us to perfectly reconstruct the measured signals from the wavelet coefficients $B_i^k(m)$, as:

$$S_i^k(\ell_n) = \sum_{m=1}^{I_k} B_i^k(m) \phi_m^k(\ell_n).$$

By doing a singular value decomposition of the matrix obtained by stacking all the coefficient vectors obtained for all the 25 data sets for each profile, we can easily identify the important independent modes of variation of the shape of each line. We represent below, for each DECA line the 10 most important (with largest energy) modes into which the bank shape is decomposed.

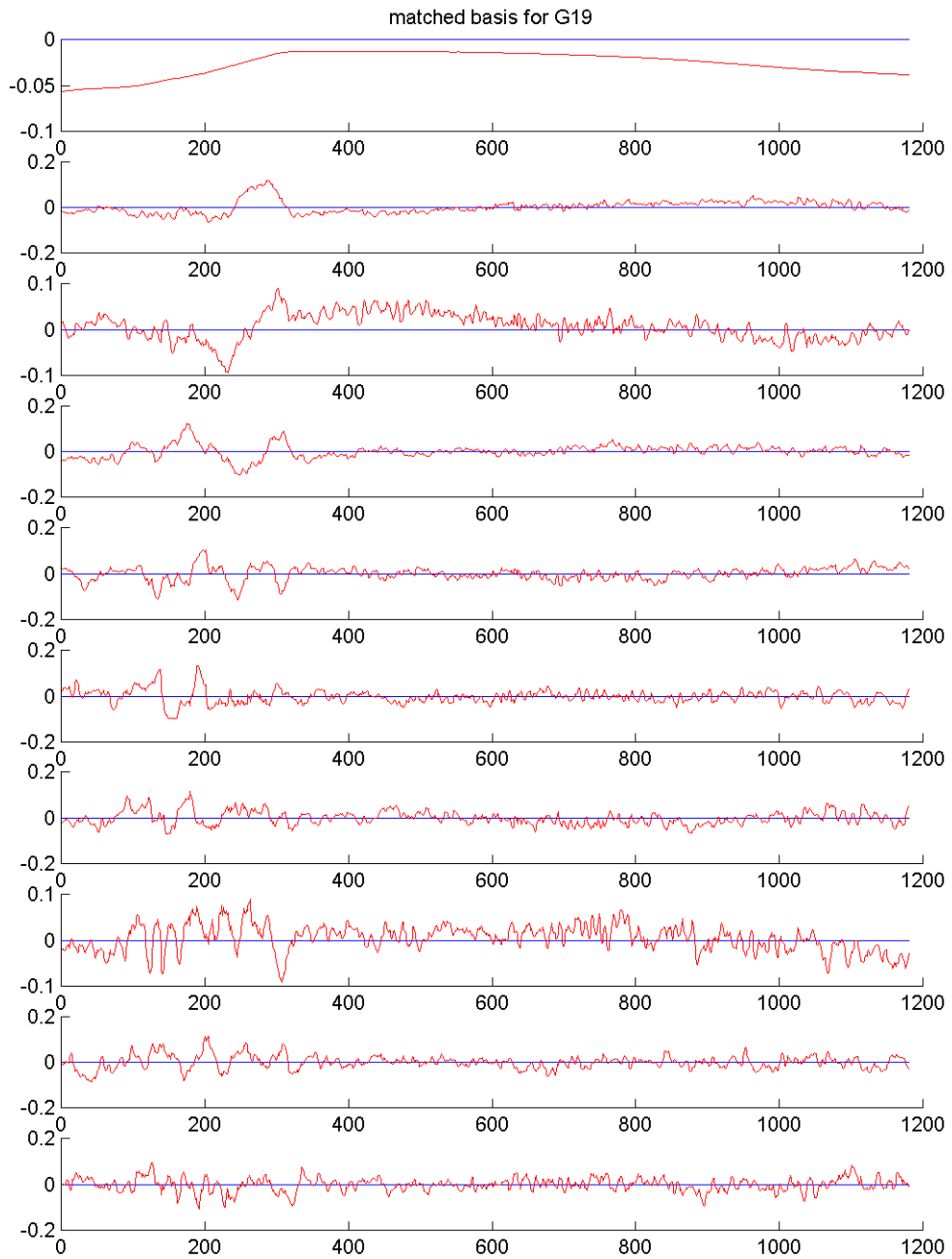


Figure B. 1: 10-dimensional basis for rG19.

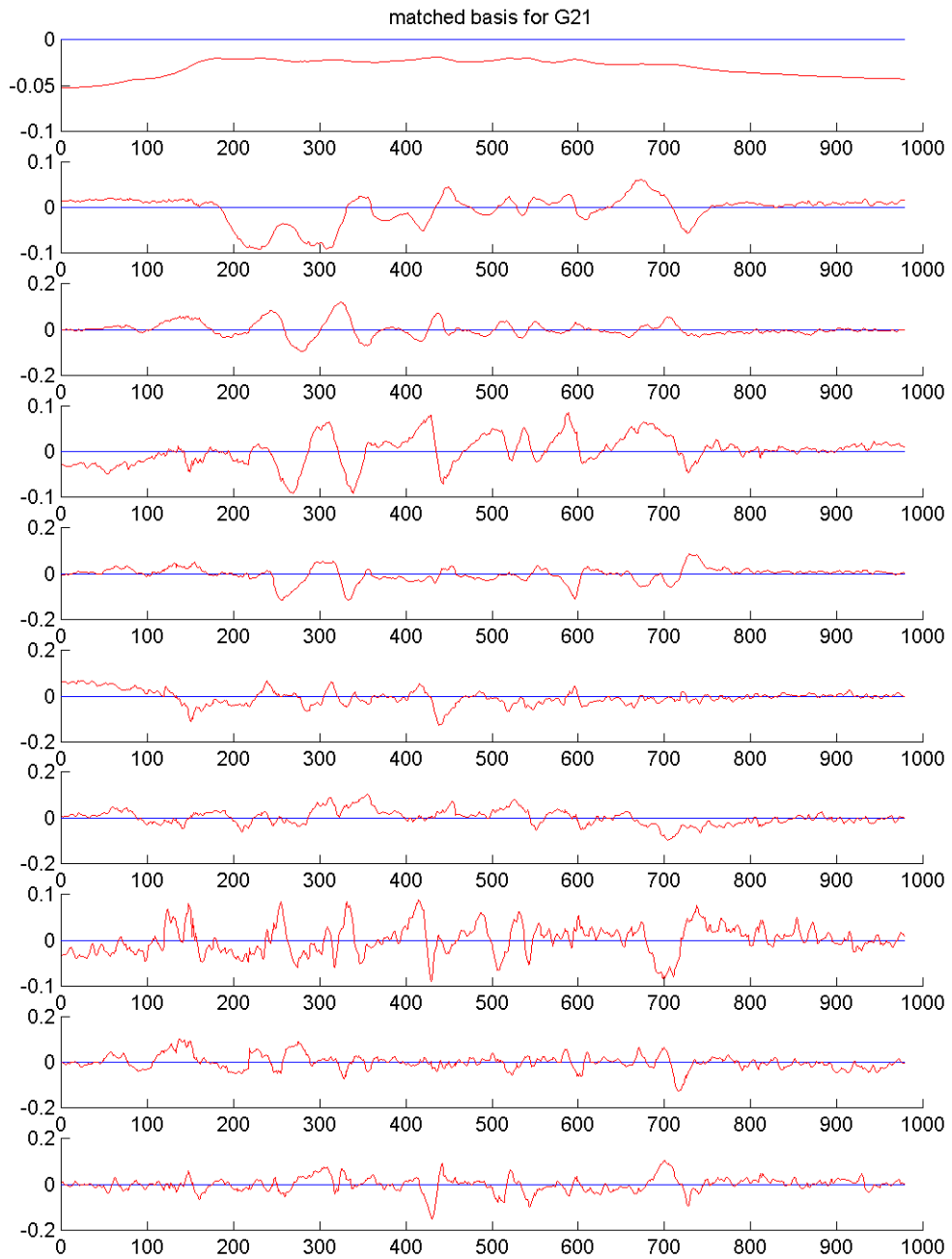


Figure B. 2: 10-dimensional basis for rG21.

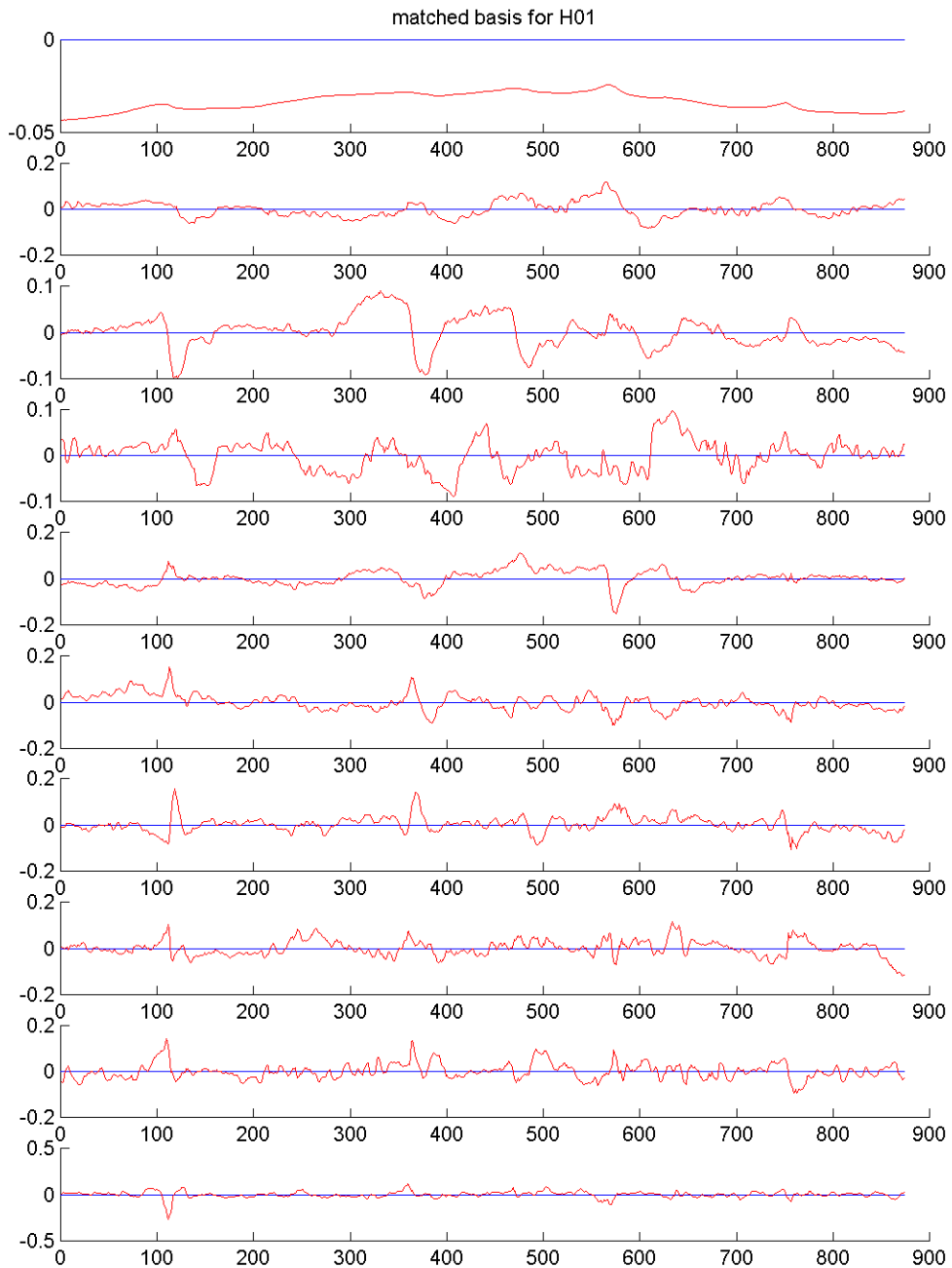


Figure B. 3: 10-dimensional basis for H01.

Note that the use of wavelets as a basis of the study enabled us to find a basis whose elements do have distinct frequency components, and with energy well localized at the important variations of the shape of the profiles.

We give below some examples of the approximation obtained with a 10-dimensional space.

Line rG19

The plot below shows (top figure) the measured (blue line) and its approximation on the linear space of the 10 basis functions obtained from the wavelet analysis (red), as well as (bottom plot) the approximation error computed over the total 25 profiles for each DECA line.

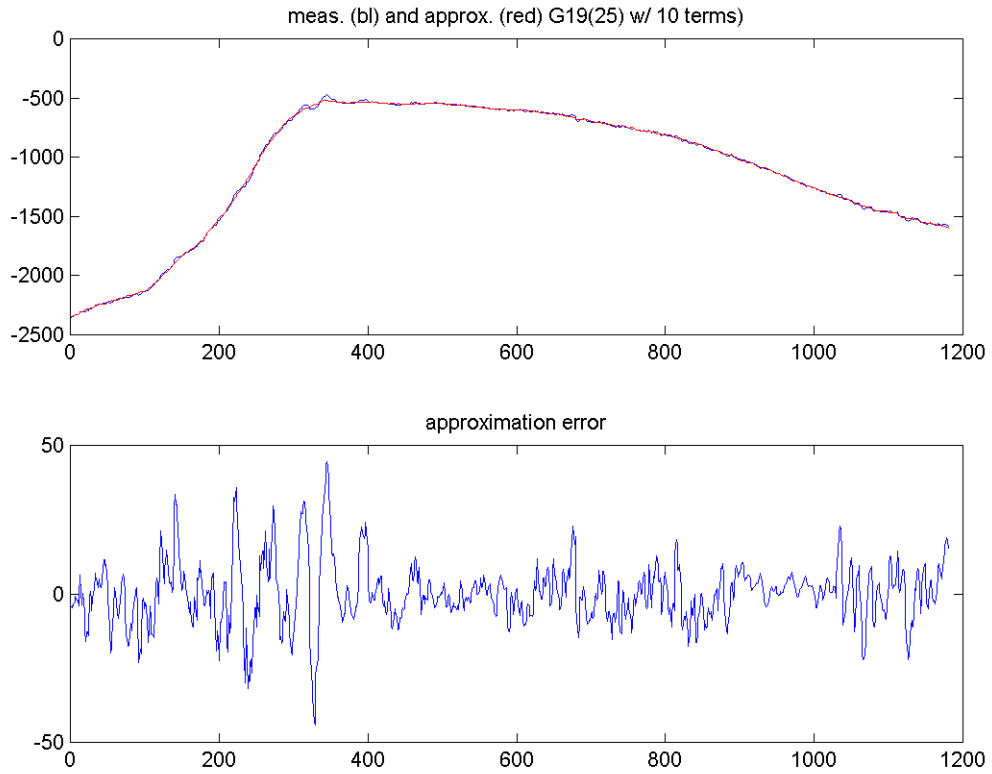


Figure B.1: G19: 10-dimensional approximation.

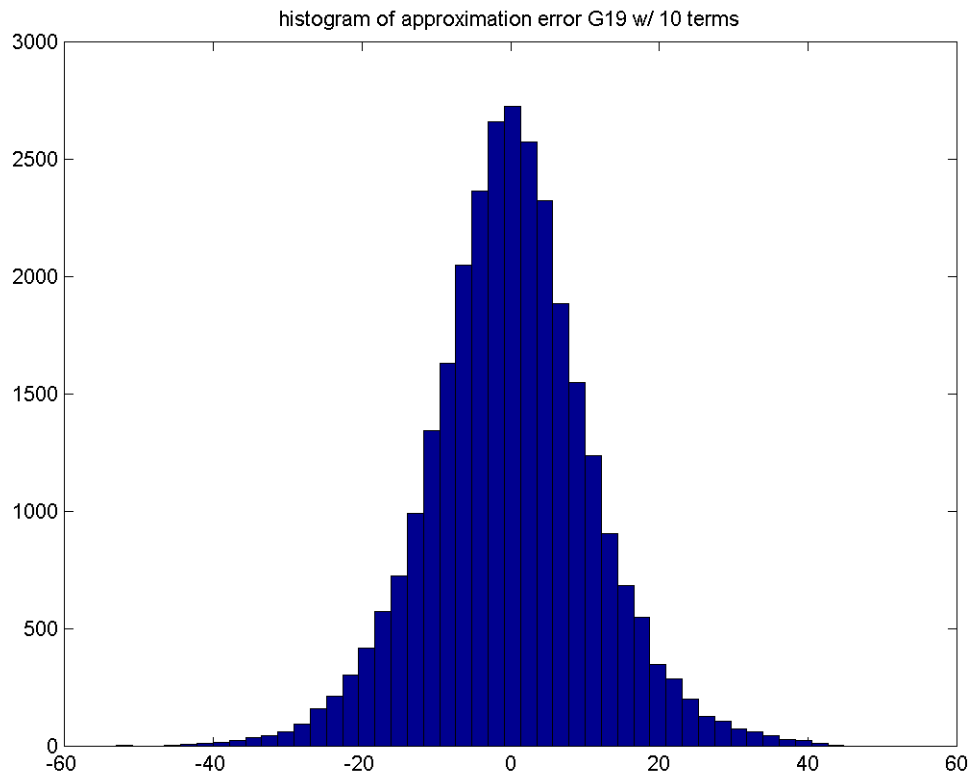
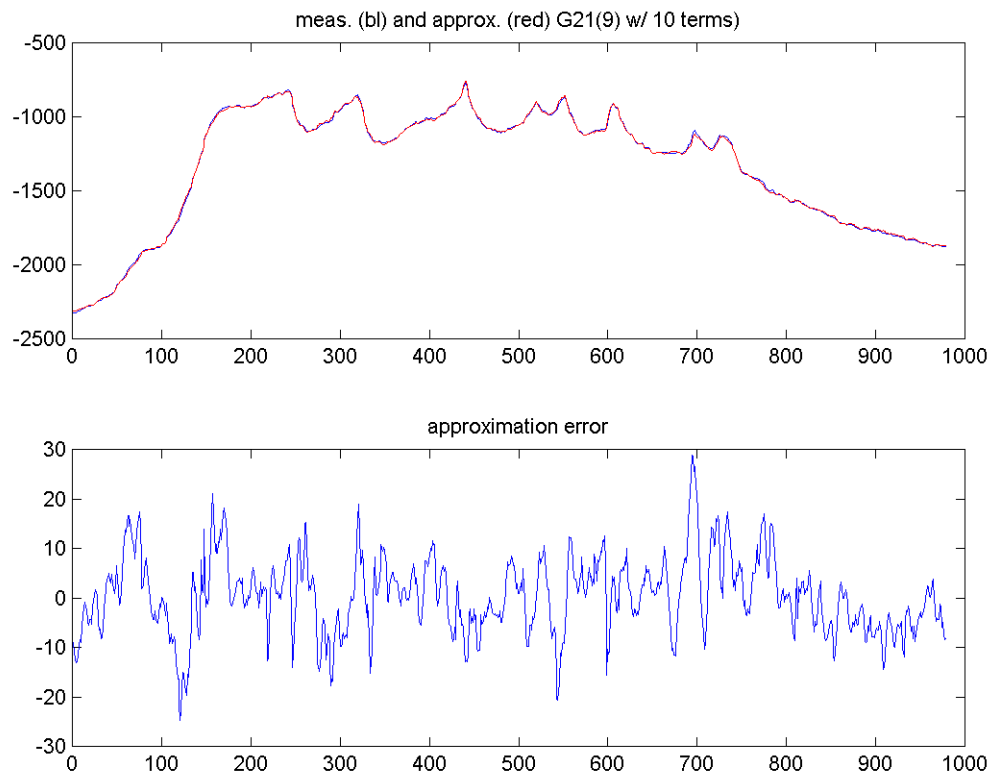


Figure B.2: histogram of approximation error for rG19.

Line rG21



FigureB.3: measured profile and its approximation (top), approximation error (bottom).

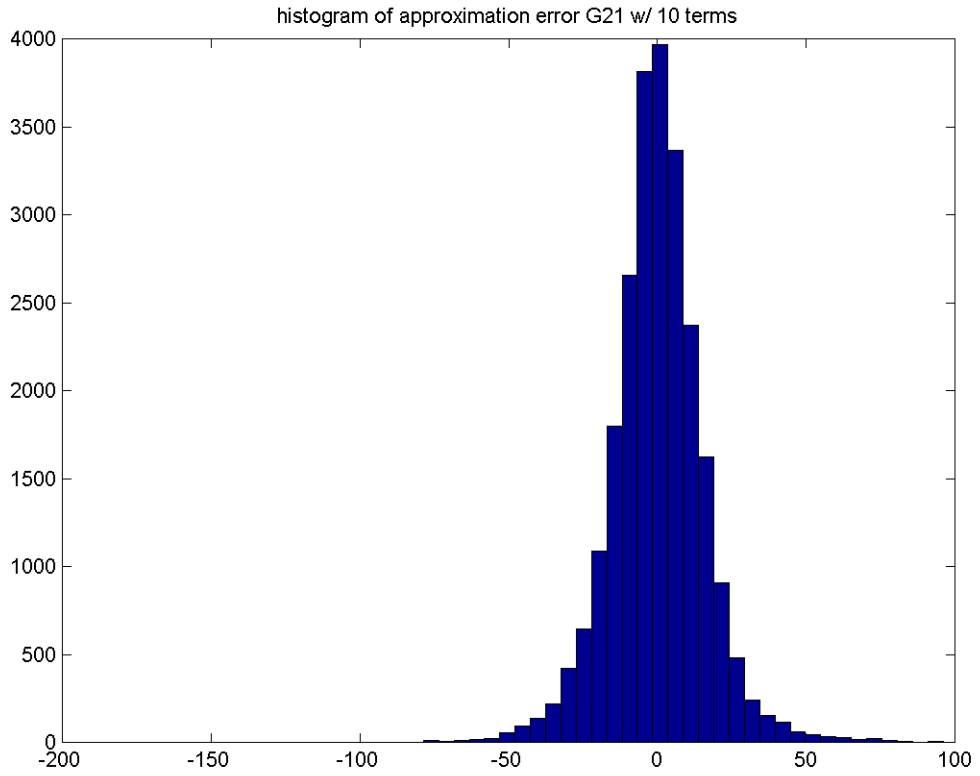


Figure B.4: histogram of the error signal on figure B.6.

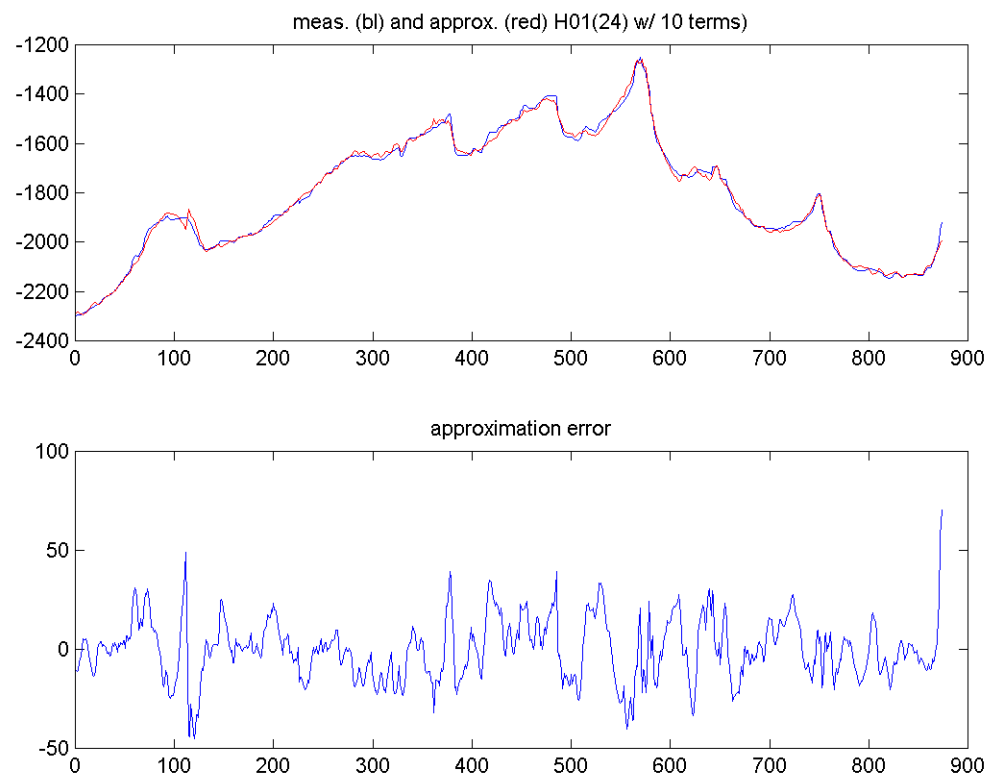
Line rH01

Figure 5: observed profile and its approximation (top) and error signal (bottom).

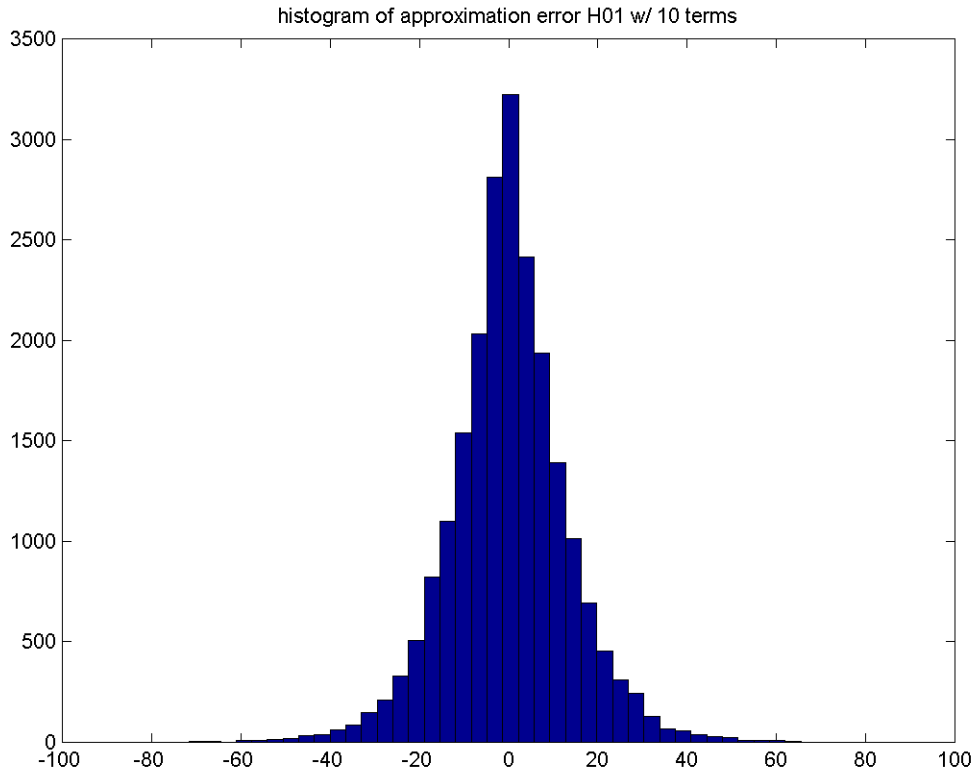


Figure B.6: histogram of the error signal on Figure B.6.

3.2 Statistical model of the bank shape

Analysis of the covariance matrix of the vector of coefficients of the representation presented in the previous section shows that it is essentially a diagonal matrix, with non-zero cross-correlation terms only between the first component and the subsequent 3 (for the basis ordering used in the plots of the previous section).

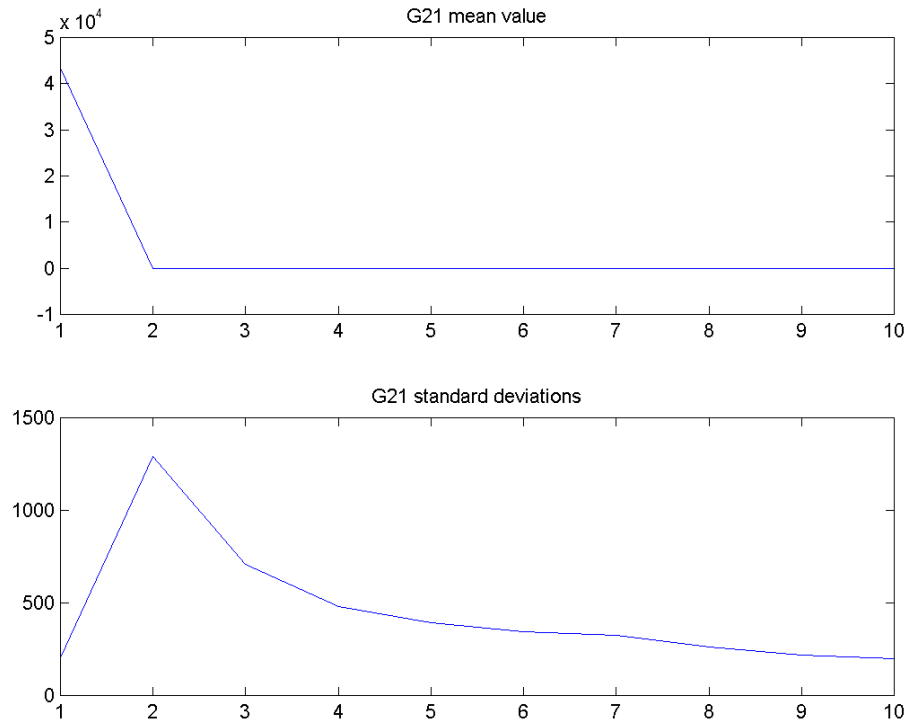


Figure B. 10: mean value and standard deviation of representation coefficients for line rG21.

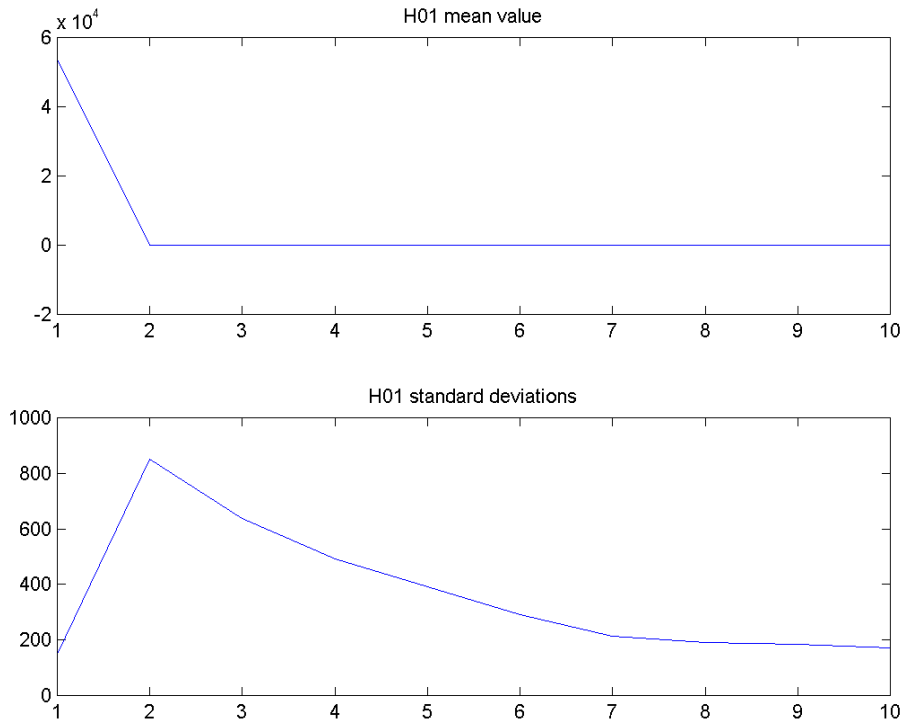


Figure B.11: mean value and standard deviation of representation coefficients for line **rH01**.

The above analysis provides a simple statistical model for the sand bank profiles along each DECA line, by assigning a Gauss distribution to the coefficients of the basis. Since the model is of small dimension (only the analysis of a larger data set would allow the determination of a complete data basis for the study), one can use this model for two purposes:

1. to produce efficient estimators of the sand bank profile integrating the knowledge contained in the learning set used in this study, allowing extrapolation of a set of measures over a small region of the profile over the entire line
2. to determine which “small regions” are the best to have a good estimate (according to the statistical model)
3. to adapt the sampling rate to the level of the actual observation noise and to the desired accuracy on the determination of the sand bank volume.

Concerning point 3 above, note that in case we are interested in determining the volume of the sand bank, using the decomposition model presented above, the volume is simply given by

$$\hat{V}^k = \sum_{i=1}^d B_i^k \int_{L_i} \hat{\phi}_i^k(\ell) d\ell$$

Once the convenient basis has been identified, we can pre-compute the integrals on the above expression, to obtain a set of coefficients

$$f_i^k = \int_{L_k} \hat{\phi}_i^k(\ell) d\ell.$$

In our case, if we compute the integral of the basis functions presented before we obtain the following plots:

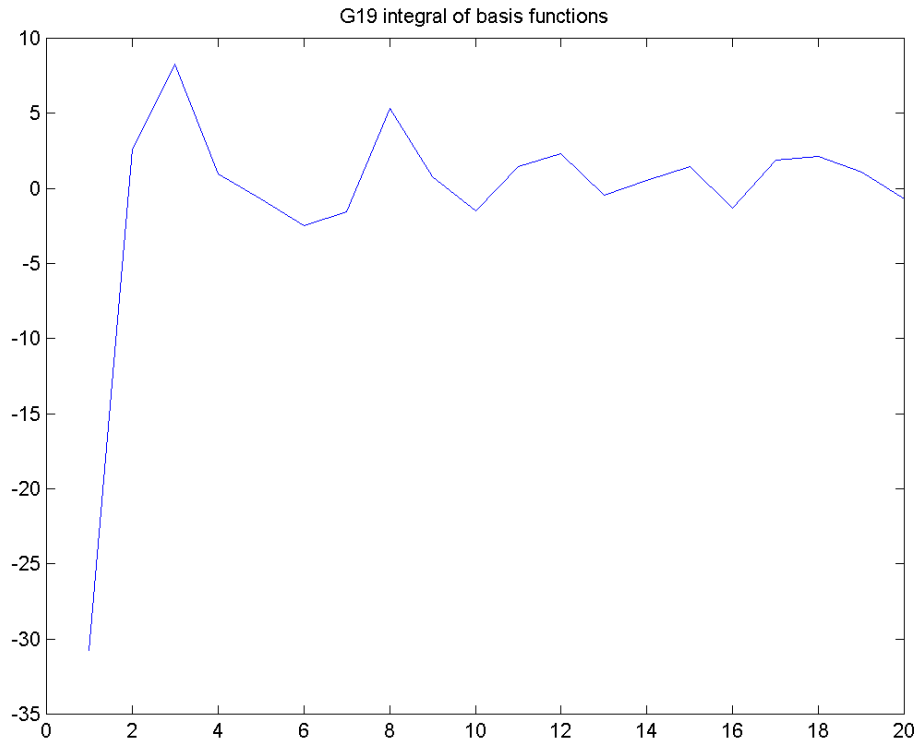


Figure 7: $\{f_i^k\}$ for line rG19.

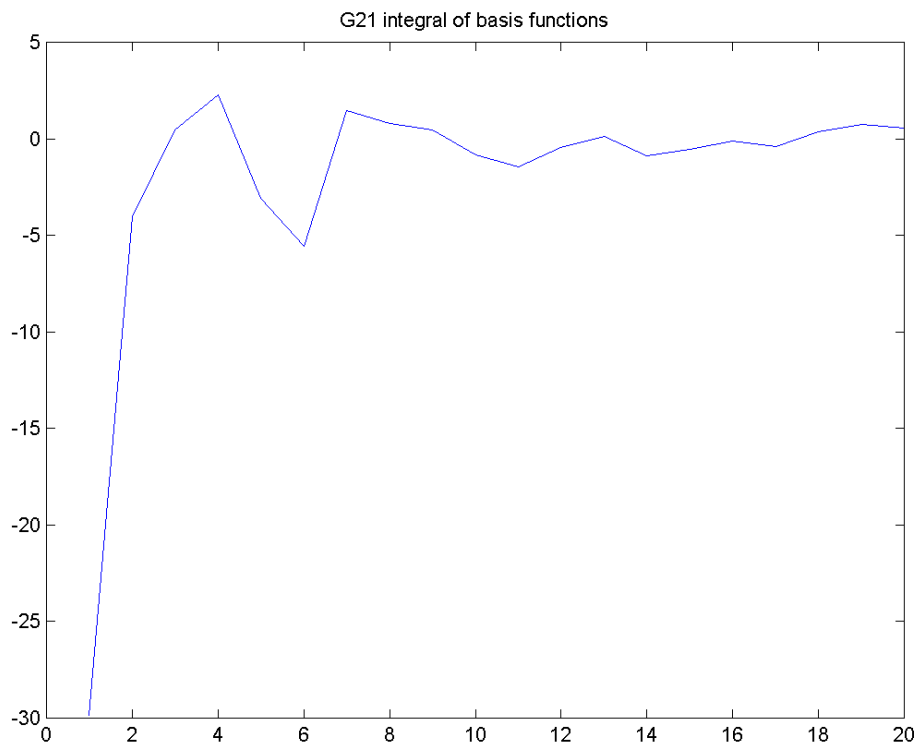


Figure 8: $\{f_i^k\}$ for line rG21.

As these plots show, the last elements of the basis have a very small contribution to the value of the volume, and the approximation of the profiles in the space spanned by the first components captures almost all of the useful information.

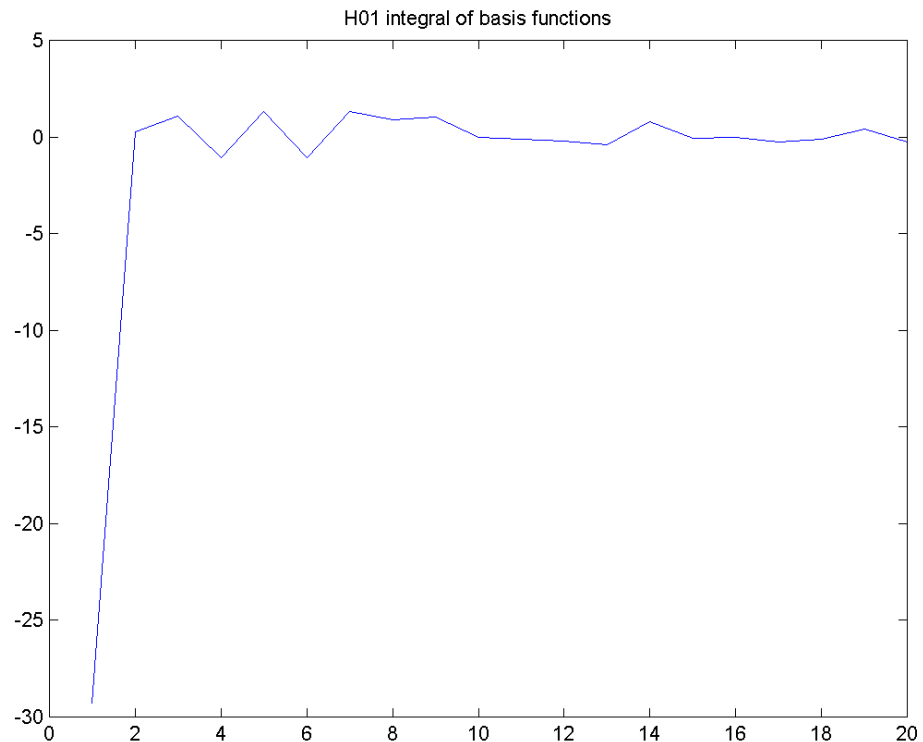


Figure 9: $\{f_i^k\}$ for line rG21.

An interesting study would be to identify a basis (eventually of higher dimension) adequate to represent **any** line across the sand bank. For this study, a larger data set, referred to the same spatial length would be required.

3.3 Structural model of the bank variations

We consider here the identification of model pertinent when one is interested in directly observing the *variations of the bank's shape*.

Let

$$\partial S_i^k = S_i^k - S_{i-1}^k$$

be the shape variation observed at survey i . Note that these differences should be normalized by the time interval between the lines. However, due to the large temporal spacing between the elements of the data set provided for analysis, we chose to do the analysis directly on the total observed variations. Better sampled data would be required to try to effectively model the shape variation as a time derivative.

Note also that we considered the restriction of the observed survey line to the intersection of the support of all surveys, and interpolated all data points to a common grid, with uniform spacing of 2 meters. In this way, the differences above are well defined.

We computed the representation of the profile's variations in a complete wavelet tree of level 9, which allows us to perfectly reconstruct the measured signals from the wavelet coefficients $C_i^k(m)$

$$\partial S_i^k(\ell_n) = \sum_{m=1}^{I_k} C_i^k(m) \phi_m^k(\ell_n).$$

Note that this decomposition factors out the time (i) and space (ℓ_n) dependencies of the time series.

Let

$$C^k = \begin{bmatrix} C_2^k(1) & \cdots & C_{N_s-1}^k(1) & C_{N_s}^k(1) \\ \vdots & & \vdots & \vdots \\ C_2^k(I_k-1) & \cdots & C_{N_s-1}^k(I_k-1) & C_{N_s}^k(I_k-1) \\ C_2^k(I_k) & \cdots & C_{N_s-1}^k(I_k) & C_{N_s}^k(I_k) \end{bmatrix}$$

be the $(N_s - 1) \times I_k$ matrix collecting all the coefficients for all observed differences. We performed a singular value decomposition of C^k :

$$C^k = U^k \Lambda_k V^k$$

Let

$$P_d = U_d^k (U_d^k)^T$$

be the projection matrix onto the linear span of the singular vectors of C^k associated to its largest d singular values. Note that C^k is at most of rank $N_s - 1$, and thus we have necessarily $d \leq 24$.

Now, let

$$\tilde{C}^k = P_d^k C^k.$$

Consider the variation reconstructed using this reduced complexity coefficient vector:

$$\partial \tilde{S}_i^k(\ell_n) = \sum_{m=1}^{I_k} \tilde{C}_i^k(m) \phi_m^k(\ell_n) = (\tilde{C}_i^k)^T \phi^k(\ell_n) = \sum_{m=1}^d \tilde{C}_i^k(m) \hat{\phi}_m^k(\ell_n)$$

where we defined

$$\tilde{C}_i^k = (U_d^k)^T C_i^k, \quad \hat{\phi}^k(\ell) = (U_d^k)^T \phi^k(\ell).$$

We see thus that considering the principal components of the coefficient vectors results in changing the wavelet basis to a reduced basis of dimension d .

We plot in the three figures below the basis determined for each DECA line (rG19, rG21 and rH01).

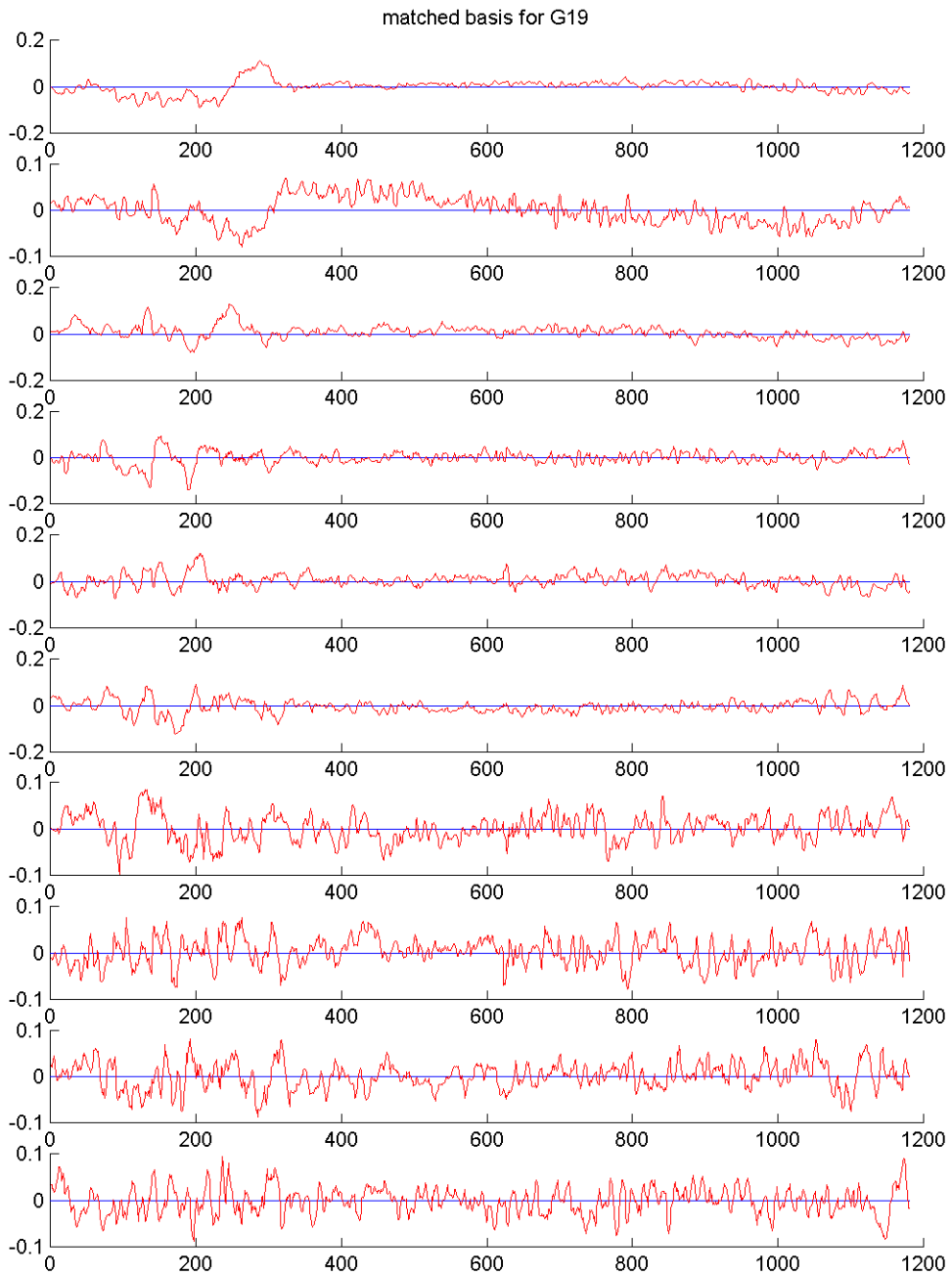


Figure B. 12: 10-dimensional basis for the variations of rG19.

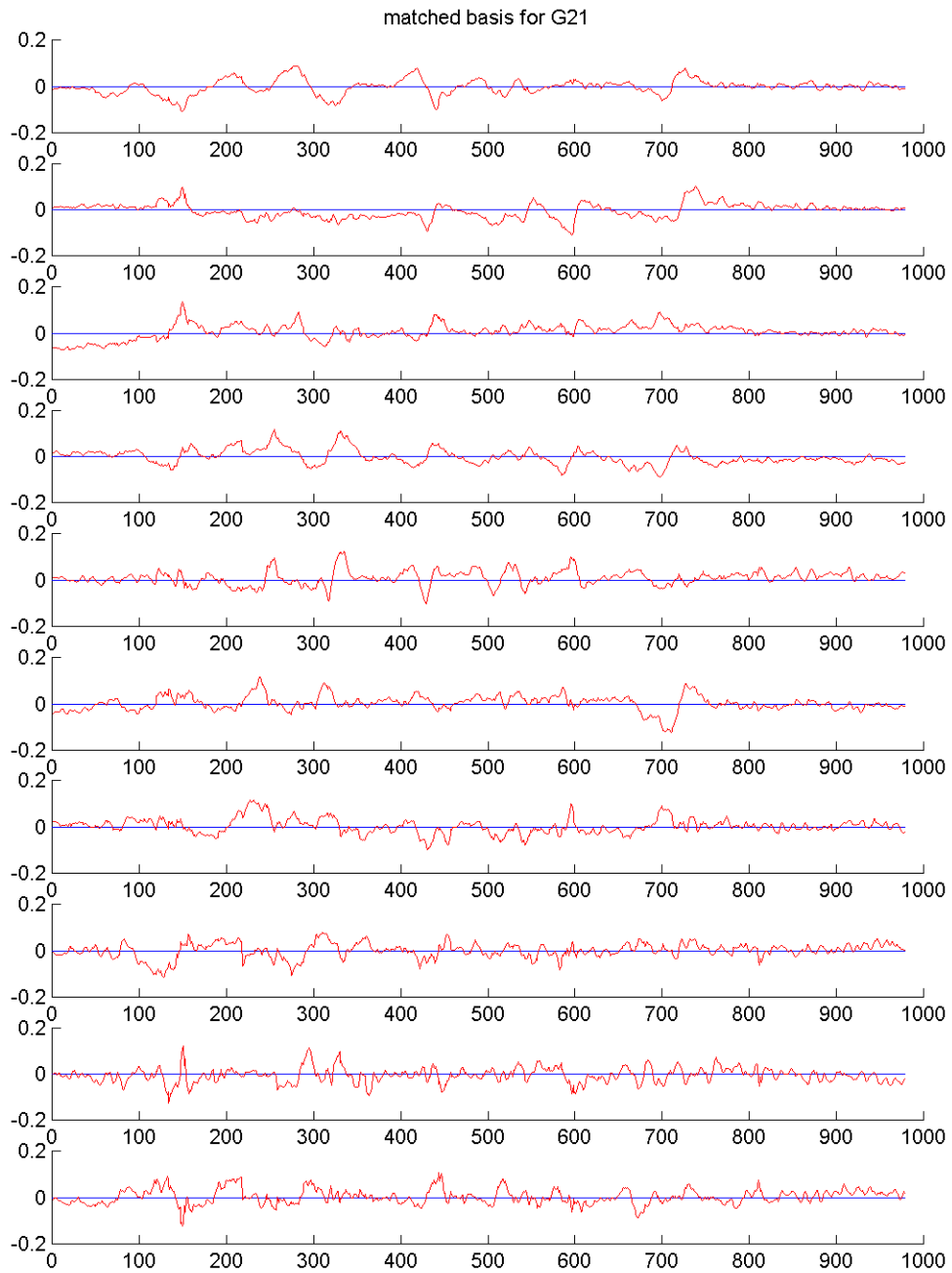


Figure B. 13: 10-dimensional basis for the variations of rG21.

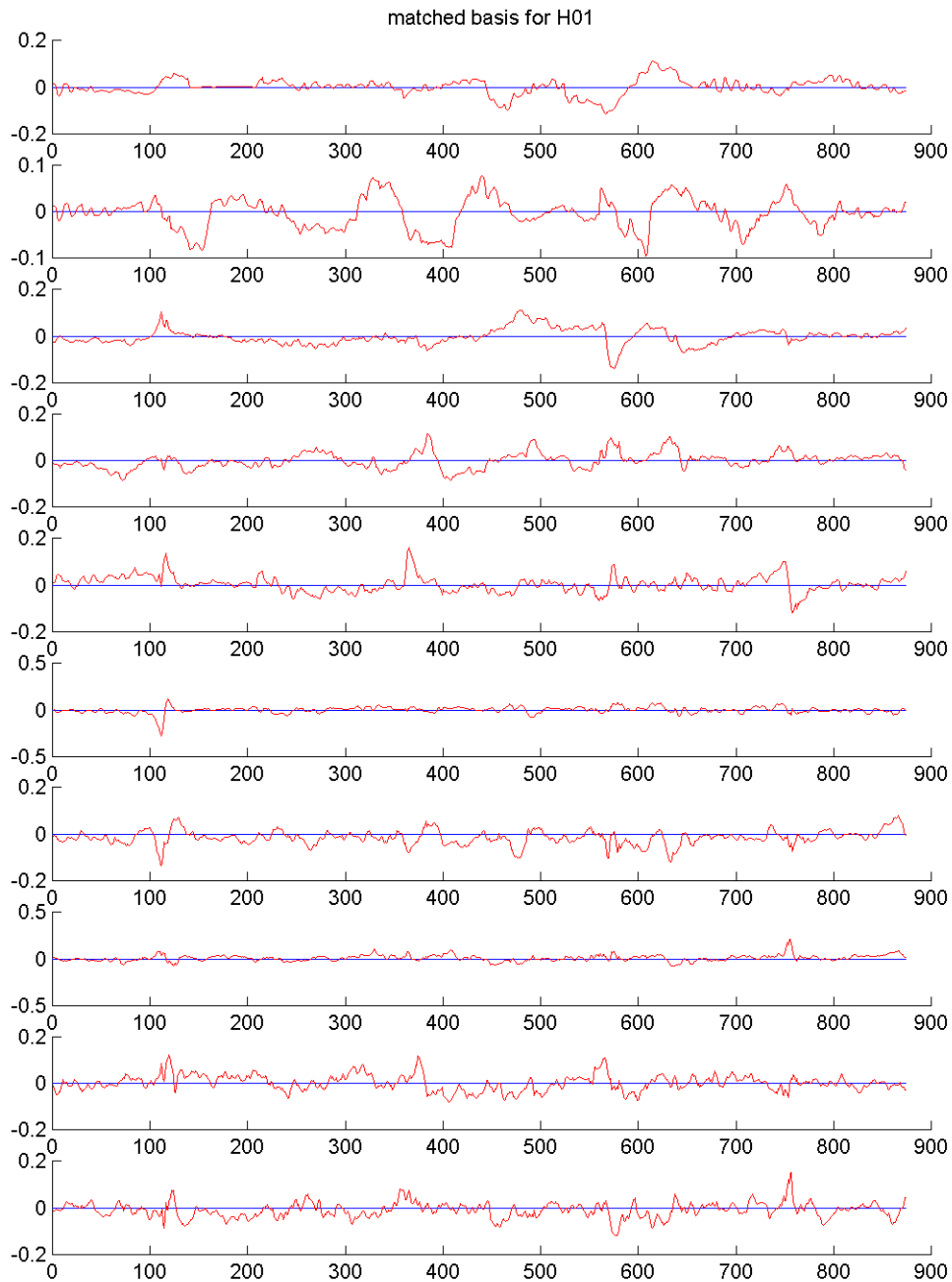


Figure B. 14: 10-dimensional basis for the variations of rH01.

We can see that the elements of the basis are localised in time and in scale. The lower frequency terms coincide well with the regions where there is effective change of the bank profile (around the dunes) and the terms of higher frequency (the last elements of the basis) are basically representing the observation noise or rapid variations which average out to a negligible modification of the shape of the bank.

For the important elements, one can also notice the fact that the adapted basis contains segments of variations of opposite signs, which represent typical modes of variation, where

sand is transported from one region to another. This has been confirmed by visual inspection of the diagrams that plot the surface obtained when the profile lines are plotted against time, as well as by the statistical analysis of the spatial correlation of the depth variations along the DECA lines.

The representation above can be used to recursively estimate the coefficients of the observed variation (if it can be directly measured) as we did before for the shape of the bank itself, by applying standard linear least-squares estimators, assuming that the observations are corrupted by Gaussian noise.

We represent below the original profiles, their reconstructions,

$$\widehat{S}_i^k = S_{i-1}^k + \partial \widehat{S}_i^k$$

and the error of the representation for values of $d=10$ and $d=15$. We also plot, for each case, the evolution of the volumes of the observed and reconstructed time series, and the histogram (over all 24 reconstructed profiles) of the approximation errors.

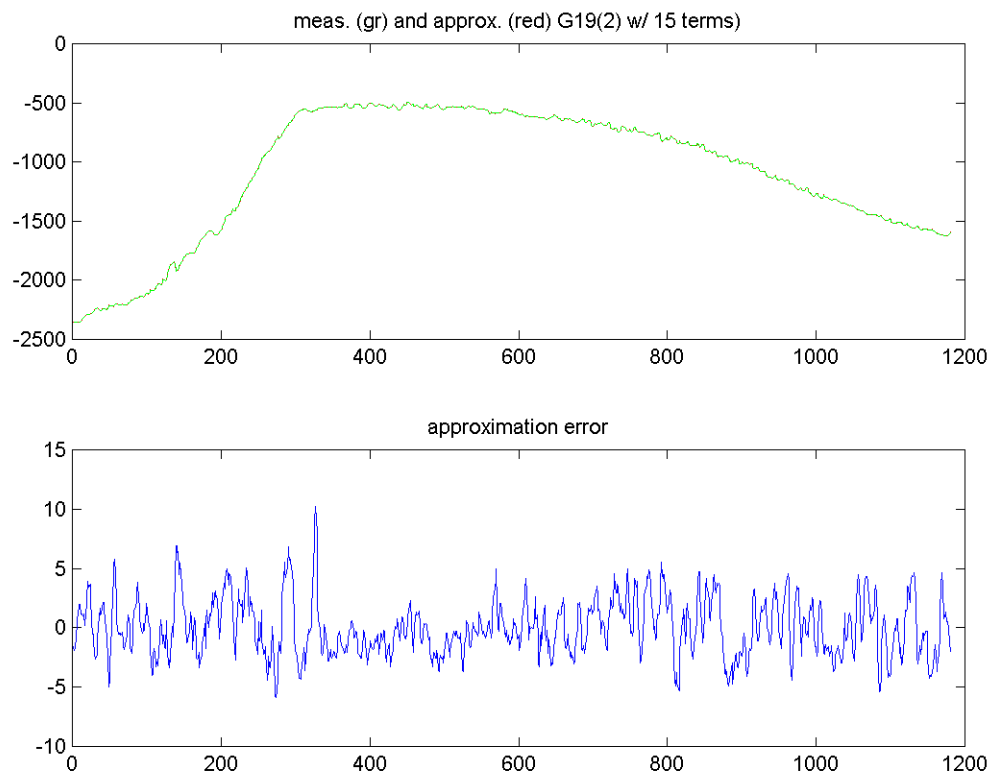
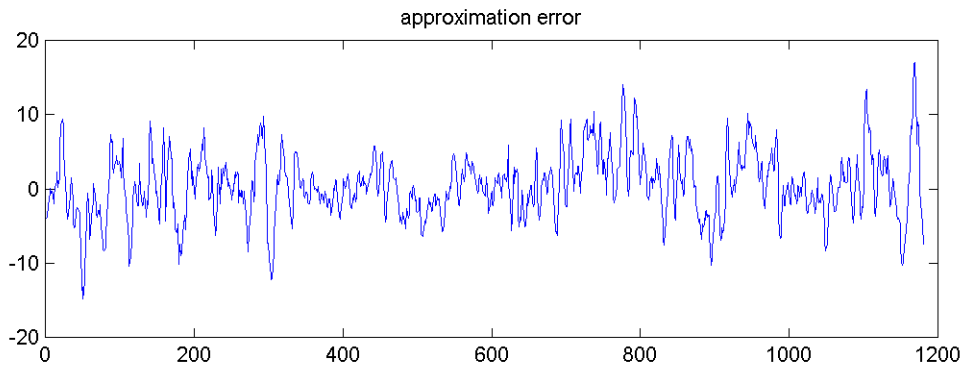
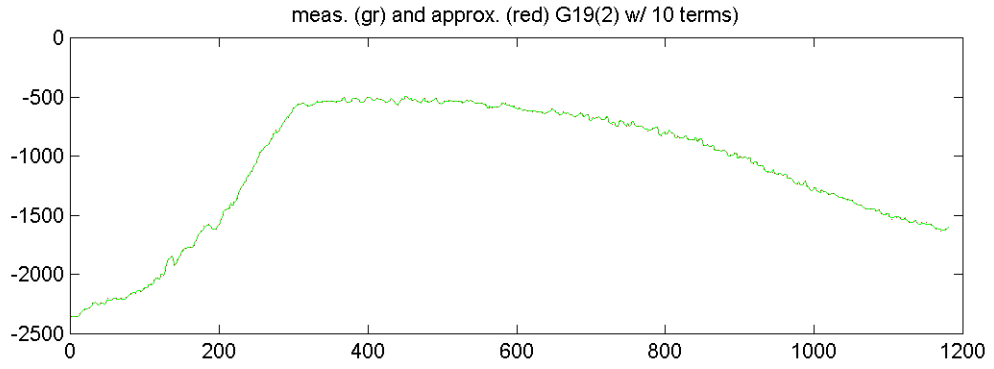
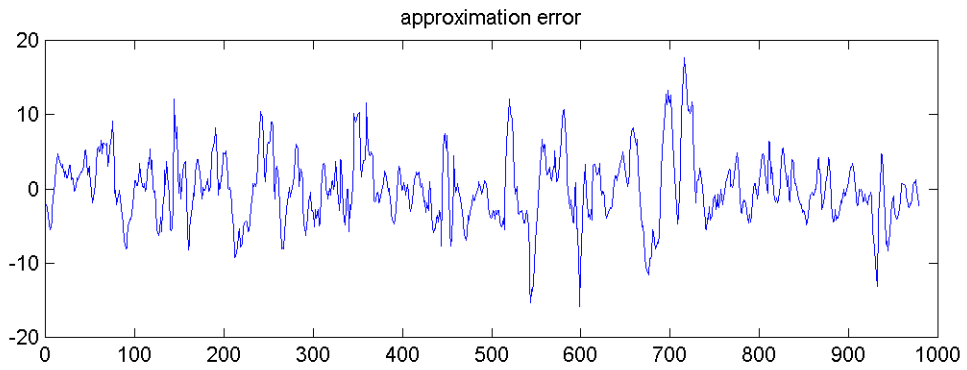
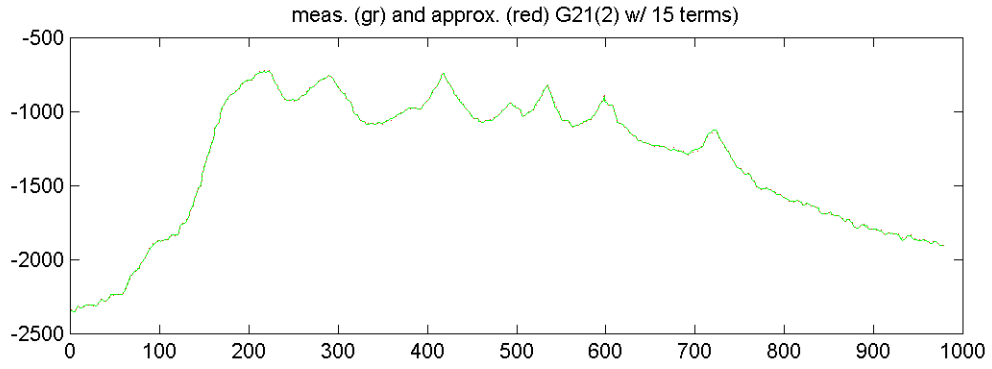
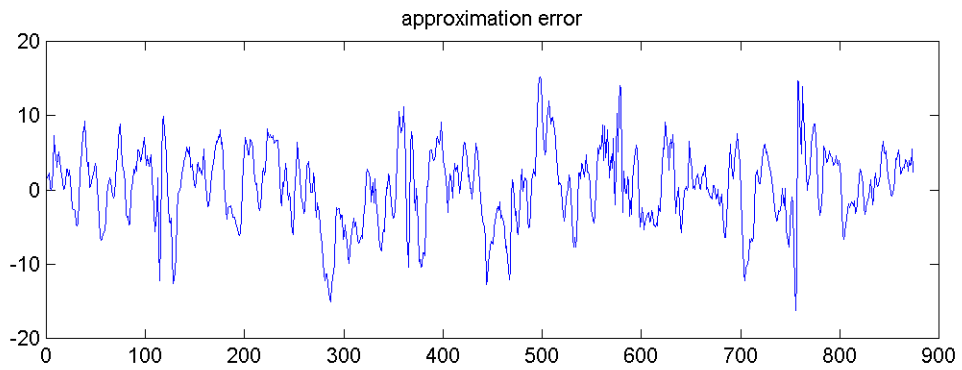
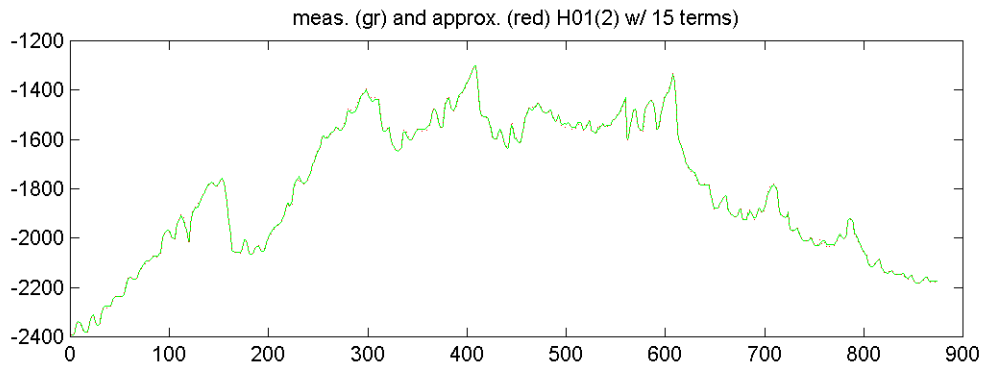
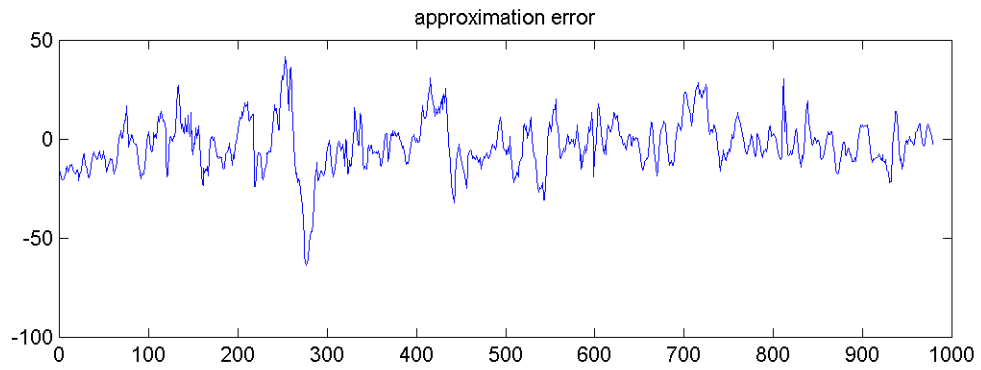
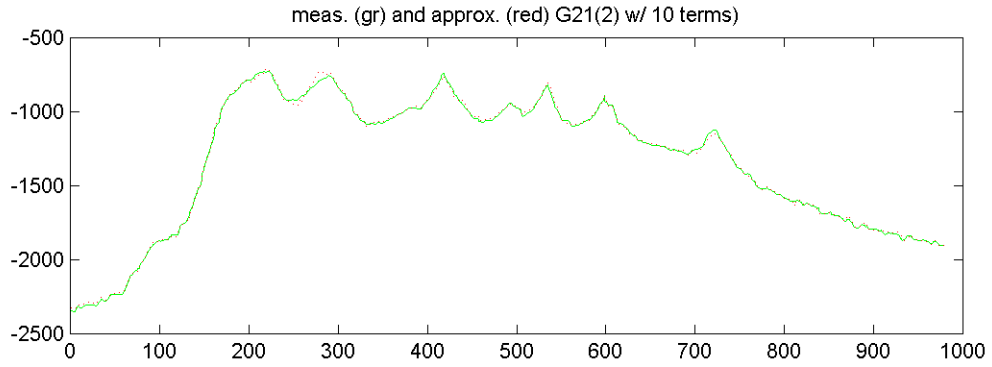
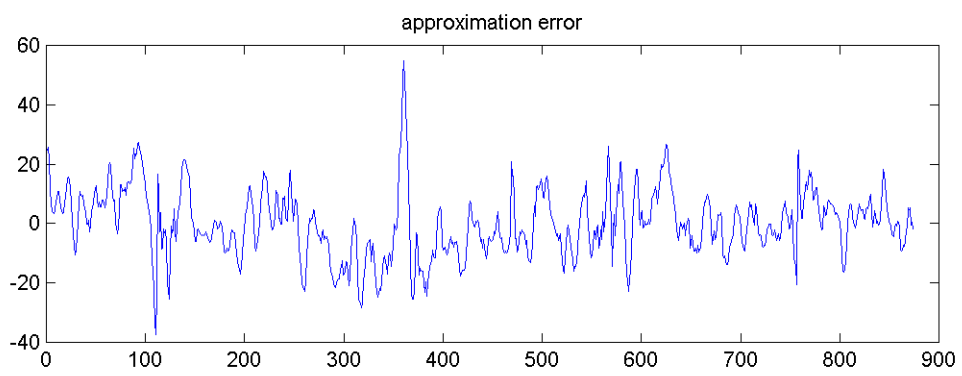
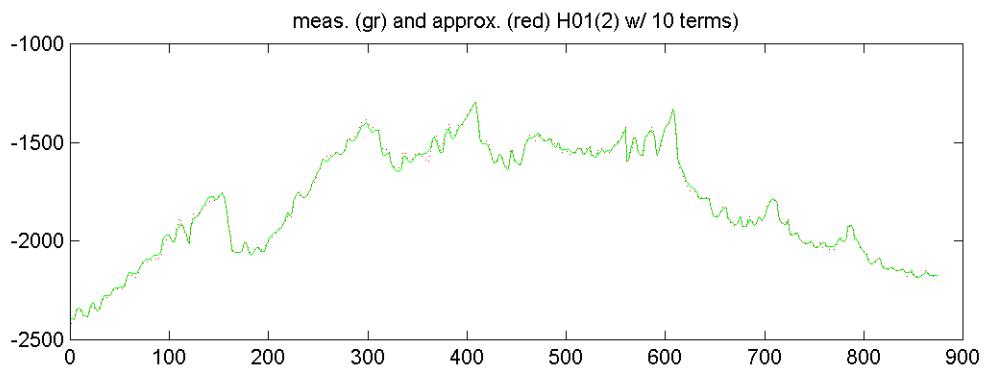


Figure B. 15: reconstructed profile from 10-dimensional approximation of difference (top) and approximation error.









Appendix 4: Image Segmentation and Recognition

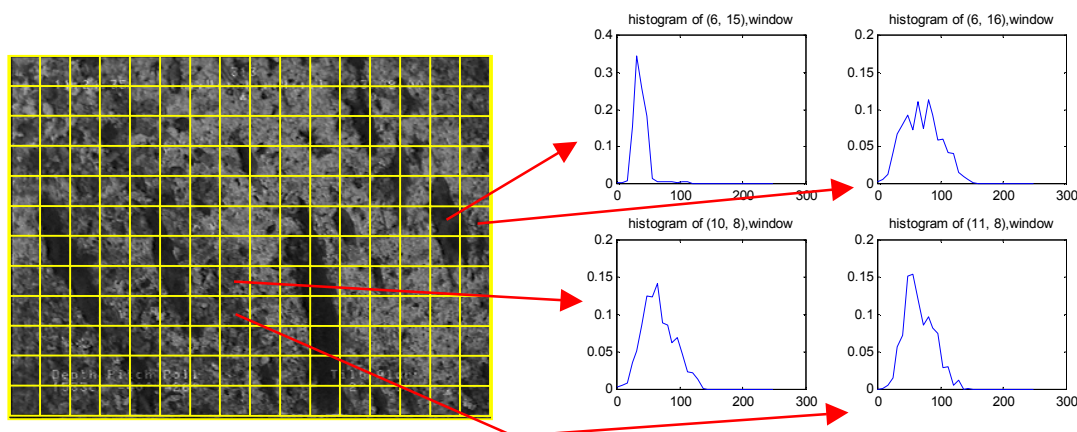
4.1 The basic image feature: the histogram

The algorithms used in the previous sections of this Chapter all rely on the same basic feature, which is extracted for each small region of the image: the histogram.

The histogram (or heuristic distribution) of a discrete random variable taking values in the finite set $\{a_i\}_{i=1}^L$ is an estimate of its probability law. Given a sample sequence $x^N = \{x_n\}_{n=1}^N$ of independent realisations of v , the histogram (or type of the sequence) is

$$h_v(a_i) = \frac{1}{N} \sum_{n=1}^N I_{x_n=a_i}, \quad i = 1, \dots, L$$

In the context of image processing, the histogram of a region of the image gives thus, the relative percentage of occurrence of the distinct grey levels. Although distinct image regions (in terms of textures) can yield sequences with the same type, in the case of the application of interest to Sumare – the discrimination between maerl and other habitats, and within the maerl class, between dead and living maerl – the distinct classes to be recognized do indeed have distinct grey levels, and different distributions around the average grey level, see next figure.



The figure above shows a real image of a sea bottom region occupied by maerl, which has been divided into small square regions, and the plots on the right display, for four of those regions, the corresponding histogram. As we can see, histograms corresponding to algae (top left histogram) are more concentrated than the other histograms, and their support confined to the dark pixel values, while regions corresponding to living maerl (two bottom histograms) span a wider number of grey levels.

Our approach to the automatic classification of an image is a mixture of *unsupervised segmentation* and *supervised classification*.

1. a first step of unsupervised segmentation *detects the homogenous regions* of the image (whose histograms are possible realisations of the same underlying probability law), and estimates the underlying probability law that better adjusts to the observed types.
2. a second step of supervised classification *assigns a label to each homogenous region* found, using a data base of histograms for each class. The construction of this data base has been detailed in Section 2.5 of this deliverable.

In this appendix, we briefly describe the two algorithms. First, we present some theoretical results from Information Theory on which they are based.

4.2 Information Theory results on the comparison of histograms

In this section we present basic results concerning the fundamental decision test on which our algorithms are based.

Consider that we are given two sequences of length n

$$\begin{aligned}x_1^{(n)} &= (x_{1_1}, \dots, x_{1_n}) \\x_2^{(n)} &= (x_{2_1}, \dots, x_{2_n})\end{aligned}$$

of independent and identically distributed discrete random variables taking values in alphabet $A = \{a_1, \dots, a_N\}$ of size $|A| = N$. The problem is to solve the **decision test**

$$\begin{aligned}H_0 : x_1^{(n)} \propto p_\mu^n \quad x_2^{(n)} \propto p_\mu^n \\H_1 : x_1^{(n)} \propto p_{\mu_1}^n \quad x_2^{(n)} \propto p_{\mu_2}^n, \mu_1 \neq \mu_2\end{aligned}$$

where the probability distributions μ, μ_1 and μ_2 are *unknown*.

We apply the *Minimum Description Length* principle [Rissanen89], which states that we should decide for the hypothesis that leads to a minimum value of the code length for coding the observed sequences:

$$\min_{i=0,1} \left\{ - \max_{\mu, \mu_1, \mu_2} \left[\ln p(x_1^{(n)}, x_2^{(n)} | H_i) \right] + L(H_i) \right\}$$

where

- The first term is the negative likelihood of the observed sequences for the best model fitting the data under hypothesis H_i
- The second term is a penalty term that corresponds to the code length required to code the parameters of the model fitted to the data under H_i .

It is known from type theory [Cover91, Dembo92] that the probability of observing a given sequence of n independent samples of a random variable distributed according to μ depends only on the type ν of the sequence, which is, by definition, the empirical estimate of its probability law (the histogram):

$$\nu_{x^{(n)}}(a_i) \equiv \frac{1}{n} \sum_{j=1}^n 1_{a_i}(x_j), i = 1, \dots, N;$$

being given by

$$P_{\mu} [x^{(n)} = x] = e^{-n[H(v_x) + D(v_x \| \mu)]},$$

where $H(p)$ is the Shannon entropy of the probability law p , and $D(p \| q)$ is the Kullback-Leibler directed divergence (also called relative entropy) between p and q :

$$H(p) = -\sum_{i=1}^N p_i \log p_i, \quad D(p \| q) = \sum_{i=1}^N p_i \log \frac{p_i}{q_i}$$

To compute the first term we must thus find

Under **hypothesis** H_0 (the two sequences come from the **same statistical model**), the best estimate of μ using the two sequences $x_1^{(n)}$ and $x_2^{(n)}$ is the histogram of the sequence of length $2n$ which is the concatenation of the two original sequences. Let v_1 and v_2 be the types of sequences $x_1^{(n)}$ and $x_2^{(n)}$, respectively. Then

$$\hat{\mu} = \frac{1}{2}(v_1 + v_2)$$

i.e., the empirical estimate for the concatenated sequence coincides with the *balanced mixture of the two types*. In this case, the first term (which actually measures the optimal code length required to code the sequence using the estimated model), is

$$-\max_{\mu} [\ln p(x_1^{(n)}, x_2^{(n)} | H_0)] = -\ln \left[e^{-n[H(v_1) + D(v_1 \| \hat{\mu}) + H(v_2) + D(v_2 \| \hat{\mu})]} \right]$$

since the two sequences are supposed statistically independent.

- Under **hypothesis** H_1 (the sequences come from **distinct models**), the best estimates (in the Maximum Likelihood sense) of the two probability laws coincide with the empirical estimates for each sequence:

$$\hat{\mu}_1 \equiv v_1, \quad \hat{\mu}_2 \equiv v_2.$$

Accordingly,

$$\begin{aligned} -\max_{\mu_1, \mu_2} [\ln p(x_1^{(n)}, x_2^{(n)} | H_1)] &= -\ln \left[e^{-n[H(v_1) + D(v_1 \| \hat{\mu}_1) + H(v_2) + D(v_2 \| \hat{\mu}_2)]} \right] = \\ &= -\ln \left[e^{-n[H(v_1) + H(v_2)]} \right] \end{aligned}$$

since the directed divergence is zero in this case.

Noting that the type of a sequence of finite length n belongs to a finite subset \mathbf{L}_n of \mathcal{Q}^n , where \mathcal{Q} is the set of rational numbers, since all its elements can be written as

$$v_i = \frac{n_i}{n}, \text{ for some } n_i \in \{0, \dots, n\},$$

we can compute a bound on the size of the set of all possible types obtained from sequences of length n :

$$|\mathbf{L}_n| \leq (n+1)^{|A|}.$$

In fact, a stricter bound can be found by noting that since their components must sum 1, the number of elements that need to be specified is just $|A| - 1$:

$$|\mathbf{L}_n| \leq (n+1)^{|A|-1}.$$

This shows that the number of parameters that are required to code an empirical estimate of a probability law defined over the discrete alphabet A obtained from a sequence of length n is at most

$$L(\nu) \leq \log(|\mathbf{L}_n|) = (|A| - 1) \log(n+1).$$

We can now establish the **code length** required to code the parameters of each model (the empirical estimates of the distribution laws):

- For **hypothesis** H_0 :

$$L(H_0) = (N-1) \log(2n-1)$$

since we estimate a single distribution defined over a set of dimension N using an effective sequence of size $2n$.

- For **hypothesis** H_1 :

$$L(H_1) = 2(N-1) \log(n+1)$$

since we estimate two distributions, each using a sequence of size n .

We can finally write the **MDL test** for this problem:

$$N[H(\mathbf{v}_1) + N(\mathbf{v}_2)] + \begin{matrix} H_0 \\ > \\ < \\ H_1 \end{matrix} N[H(\mathbf{v}_1) + H(\mathbf{v}_2) + D(\mathbf{v}_1 \|\hat{\mu}) + D(\mathbf{v}_2 \|\hat{\mu})] +$$

$$2(N-1) \log(n+1) \qquad (N-1) \log(2n+1)$$

which is equivalent to

$$\frac{(N-1)}{N} (2 \log(n+1) - \log(2n+1)) \begin{matrix} H_0 \\ > \\ < \\ H_1 \end{matrix} D(\mathbf{v}_1 \|\hat{\mu}) + D(\mathbf{v}_2 \|\hat{\mu}) = L_{\mathbf{v}_1 | \mathbf{v}_2}$$

and we see that the optimal test compares the **distance between each individual type and their balanced mixture** to a threshold that depends on the number of parameters N (the size of the input alphabet) and on the number n of data points. If the distance is smaller than this threshold the

decision is that the sequences come from the same model, otherwise, that they come from different models.

Note that the optimal test does not directly compares the two types v_1 and v_2 , as one could guess at a first time. In fact, using the convexity of the Kullback divergence [Cover91]:

$$D(v \| \alpha p + (1 - \alpha)q) \leq \alpha D(v \| p) + (1 - \alpha)D(v \| q)$$

we can write, since in our case $\alpha = 0.5$, that

$$\begin{aligned} D(v_1 \| \hat{\mu}) + D(v_2 \| \hat{\mu}) &\leq 0.5 [D(v_1 \| v_1) + D(v_1 \| v_2) + D(v_2 \| v_1) + D(v_2 \| v_2)] \\ &= 0.5 [D(v_1 \| v_2) + D(v_2 \| v_1)] \end{aligned}$$

We make several notes here:

1. The test defined in terms of the types of the two original sequences are defined only if they are mutually absolutely continuous with respect to each other. This is in general not the case when dealing with small samples defined over a reasonably sized alphabet. The optimal algorithm, on the other hand, is always well defined, since the types are necessarily absolutely continuous with respect to the mixture measure.
2. Even when the previous problem does not occur (we can set bounds for its probability), utilisation of the average of the two measures leads to a test that tends to favour hypothesis H_1 (distinct models).
3. Another test in terms of the maximum value of the directed divergences can also be obtained, since the average of two values is always smaller than their maximum. This test has the problems pointed above.

4.3 Unsupervised segmentation algorithm

We have designed two unsupervised segmentation algorithms based on the results presented in the previous section.

Anisotropic diffusion algorithm

The first is an anisotropic diffusion algorithm. It starts by assigning to each window W_{ij} of the image the type of the original image pixels in that region:

$$h_{ij}^0(a_i) = \frac{1}{\#W_{ij}} \sum_{n \in W_{ij}} 1_{i_n} = a_i$$

For each window W_{pq} in a neighbourhood $N(W_{ij})$ of the window W_{ij} , the test described in the previous section is computed, and the histogram updated by replacing the current histogram by a mixture of itself with the neighbour histograms

$$h_{ij}^k = \alpha h_{ij}^{k-1} + \sum_{pq \in N(W_{ij})} \beta(p, q; i, j) h_{pq}^{k-1}$$

where the weights satisfy the following conditions

1. $\alpha > 0, \beta(p, q; i, j) \geq 0$
2. $\alpha + \sum_{pq \in N(W_{ij})} \beta(p, q; i, j) = 1$
3. $\frac{d\beta(p, q; i, j)}{dL_{p, q| i, j}} < 0$

where $L_{p, q| i, j}$ is the optimal test to decide if the two types come from the same or from distinct probability laws presented before.

As defined above, this algorithm will converge to a situation where the image is divided into regions R_n that share a common histogram:

$$\lim_{k \rightarrow \infty} h_{ij}^k = h^{c_n}, W_{ij} \in R_n.$$

Different annealing schedules for the mixing coefficients have been studied, and good results have been obtained with exponential annealing. Details can be found in [Tenas2001].

Lloyds algorithm in distribution space

The other algorithm is conceptually equivalent for the Lloyds algorithm [Lloyd82] (from vector quantization theory) that iteratively builds a partition of the input space, which is a centroidal tessellation (the generators of all Voronoi cells coincide with the centroids of the cells). In our case, the input space is the space of types for finite length sequences of *iid* realisations of the same random variable, with a geometry induced by the Kullback divergence while for classical applications of the Lloyds algorithm, the input space is the Euclidean space of dimension n .

The Lloyds algorithm iteratively alternates between finding the optimal tessellation for a set of generators (by using minimum Euclidean distance), and re-computing new generators as the centroids of the current tessellation. Our version of the Lloyds algorithm partitions the input space according to the Kullback distance of all points with respect to the current centroids, and re-computes new centroids as the mixture of all points assigned to the same cell.

When a convenient centroidal tessellation has been found, the size of the partition is doubled, by randomly generating two new generators from each existing one. In the Lloyds algorithm, this is done by adding random perturbations of opposite sign to each existing centroid. In our case, we must make sure that the resulting perturbations are members of the L -1 dimensional simplex of valid probability laws.

One major limitation of the Lloyds algorithm is the fact that the number of classes (of cells) must be known in advance, and, moreover, the algorithm is efficient only for powers of two. Our version of the Lloyds algorithm, designed to work in the type space, overcomes this limitation by using another result from Information theory [Dembo92], which states that, asymptotically, the distribution of the Kullback divergence of the type of a sequence of length n to the true distribution follows an exponential law:

$$\Pr\{D(h_v(x^n) \parallel \mu) > d\} \leq e^{-nd}$$

Once a stable partition is found, the purity of each class is tested by evaluating the departure from exponentiality of the distribution of the members of the class to the centroid of the class. Compared to the previous one, this algorithm presents the important advantage of being much faster (compatible with real-time segmentation for guidance of the vehicle).

4.4 Classification algorithm

The algorithms presented in the previous section learn the probability distributions that better adjust to the regions present in one given image, but cannot tell us which species they correspond to. As it is explained in section 2.7 of the deliverable, a data basis of types associated to the main classes of interest in the study have been learned

$$\text{Class } C_m \leftrightarrow \{h_1^m, h_2^m, \dots, h_{N_k}^m\} \quad k = 1, \dots, K$$

The optimal classification algorithm assigns, to window W_{ij} , to which one of the previous unsupervised algorithms has made correspond the type h_{ij}^∞ , the class that minimizes the Kullback divergence between h_{ij}^∞ and the members of the data basis:

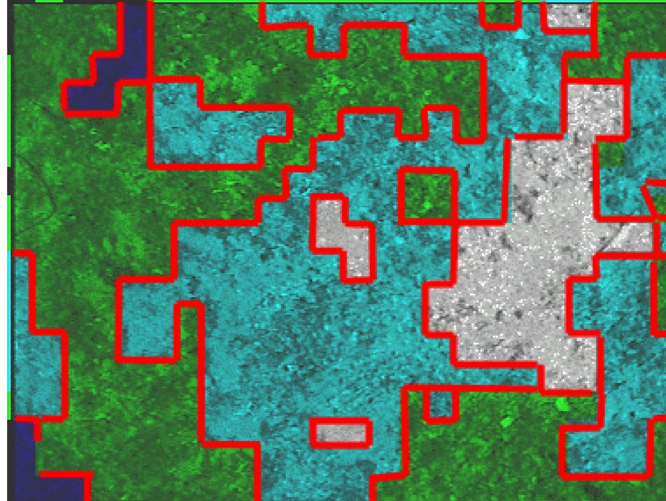
$$W_{ij} \leftrightarrow \text{class } C_k \leftrightarrow \min_n D(h_{ij}^\infty \parallel h_n^k) < \left\{ D(h_{ij}^\infty \parallel h_p^\ell) \right\}_{p=1}^{N_\ell}, \quad \ell \neq k$$

This is the optimal (minimum error probability) test for the class. We make some notes:

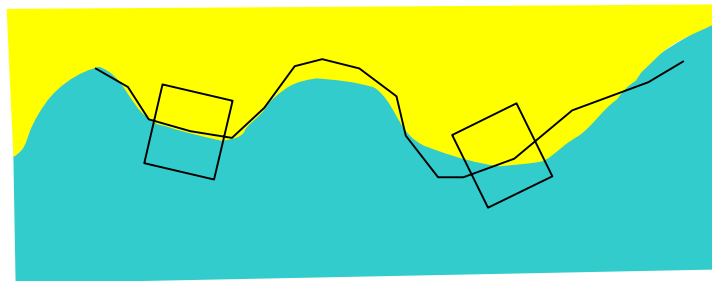
1. The test is highly dependent on the available data basis of histograms for the different classes that must be recognised. It is important that typical, as well as not-so-typical examples be included in the test, to accurately delineate its decision boundaries.
2. The test assigns a class to each class independently of the histograms that may be classified in the same image for different classes. There is clearly room for improvement of this test by considering, instead of independent models for each class, models for the joint appearance of all classes. In particular, this might alleviate the some problems encountered with the automatic contrast adjustment of the camera, which induces spurious variations on the grey level of the images.

4.5 Contour tuning

The image segmentation algorithms that we presented above are based on a partitioning of the original image into small square regions, and consequently they produce "staircase" contours between the identified regions, see example below



There may be interest in adjusting these contours to the boundaries of the regions present in the image. Based on the same tools that the previous algorithms, we defined an iterative algorithm that gradually deforms the contour by applying to each point a force that tends to centre the contour on the boundary. The basic idea behind this algorithm is the following: *“if we centre a square window on a true contour point, aligned with the tangent to the contour at that point, the type associated to the pixels inside the window should be a balanced mixture of the histograms associated to the classes on each side.”* The figure below illustrates this idea:



The first window (on the left) is well centered on the contour, having an equal percentage of each class, while the second window (on the right) has more of the yellow class than of the blue one.

Ideally, we should determine, at each point of the contour, the type of the sequence in the window centred at that point and aligned with the tangent to the contour, \hat{h}_x and estimate from it the the mixing parameter in the following model:

$$\hat{h}_x = \alpha h_1 + (1 - \alpha) h_2$$

where h_1 and h_2 are the histograms associated (by the previous algorithms) to the classes on the left and right side of the contour. In fact, we work over a finite number of equally distributed points on the contour.

Once an estimate of α is available ($\hat{\alpha}$), we apply at point x of the contour a force that is normal to the curve at that point, and whose sign and norm depend on the value of $\hat{\alpha}$:

$f_x = \chi(\alpha - 0.5)$, where we assumed that positive forces act toward the right side of the contour.

Estimation of α requires minimization of a non-linear function, which basically involves the Kullback-Liebler divergence between \hat{h}_x and the mixture of the two classes' histograms with parameter α . We found sufficient to compute the likelihood of the values $\alpha \in \{0, 0.25, 0.5, 0.75\}$ and apply either a zero force (when the value 0.5 is the most probable) or a force of constant strength on one of the sizes, depending on which of the values 0.25 and 0.75 is the most probable.

The algorithm works iteratively, by uniformly sampling the current contour, deforming it according to the forces acting upon it at each point, and smoothing the deformed points using splines (which act as a regularisation term, identical to the internal forces of deformable snakes in image processing).

Appendix 5: Estimation of Statistical Spatial Model

In this appendix we briefly outline the mathematic tools on which rely our characterisation of the spatial distribution of the maerl fields. The material below is taken from [Rolfes2001].

Introduction

Random Closed Set (RCS) models are a mathematical model that has been widely used in order to analyze random patterns. We recall that a RCS is a doubly stochastic process. A first process (point process) determines the locations of the objects and a second process (shape process) the morphological characteristics of the objects placed at each location.

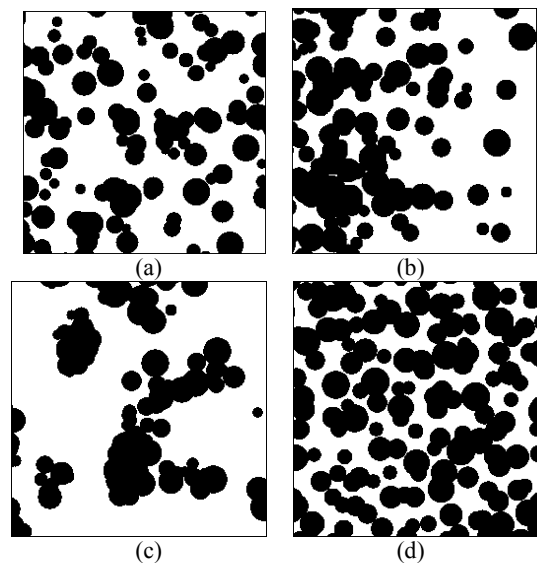


Figure 1: Example of RCS models: (a) isotropic Boolean model, (b) anisotropic Boolean model, (c) clustered distribution and (d) regular distribution of the grains.

We consider that these two processes are independent, which is not always the case for natural environments. Under the assumption of independency we can construct a family of RCS model types $\ell = \{\ell_i\}_{i=1}^N$ as a set of pairs of one point process model and one shape process model. Examples of point processes are: Poisson point process, regular distribution or clustered distribution (see Figure 1). A particular model is obtained by a model parameter $\theta_i = (\lambda_i, \gamma_i)$, where λ_i determines the point process and γ_i the shape process, means the measure defined on the shape space. (In the sequel we denote by θ_i the RCS model type and the particular model parameter). The shape process is chosen in order to restrict the possible shapes to simple basic shapes, such as line-segments of random length and orientation or compact discs of random radius. This is an approximation of generally very complex shaped objects, which are impossible to model by low dimensional parameter vectors.

Direct estimation of the spatial distribution (count measure) and the morphological characteristics is impossible. However, it is known that knowledge of the RCS model is equivalent to knowledge of the hitting capacities of the random field for all compact sets. The hitting capacity is defined as the probability that the intersection of the random field with a compact set $K \in \mathbf{K}$ is non-empty:

$$T_{\Xi}(K) = P(\Xi \cap K \neq \emptyset).$$

Estimation of model parameters

While we are not able to directly estimate the model parameter, we can obtain an estimate of the hitting capacities from the segmented images, under the assumption that the random field is locally isotropic which implies that $T_{\Xi}(K) = T_{\Xi}(K+p)$. The equivalence between the hitting capacities and the random closed set model mentioned above can be exploited to estimate the model parameters. It is thus more convenient to write $T_{\theta_i}(K) := T_{\Xi}(K)$, knowing that the model type is ℓ_i . For Boolean models (the point process is a Poisson point process and the grains are i.i.d. in the workspace) the hitting capacity can be written as a function of the compact set K shape and the model parameter θ_i [Stoyan95]:

$$T_{\theta}(K) = 1 - \exp(-\lambda E\{\nu(\Xi_0 \oplus \tilde{K})\})$$

where \oplus is the Minkowski-addition $A \oplus B = \{a+b, \forall a \in A, b \in B\}$, $E(\cdot)$ is the statistical expectation operator with respect to the measure of the shape process, $\tilde{K} = \{-x, x \in K\}$, $\nu(\cdot)$ is the Lebesgue-measure and Ξ_0 is a random shape. The basic requirement for our approach to the modelization of natural scenes and to mobile robot navigation is the ability to determine the hitting capacities as a function of the RCS model type and parameters. For the moment we are restricted to the consideration of Boolean models. In order to identify more complex models, such as regular or clustered distributions of the objects, additional efforts need to be spent.

Of course we cannot observe the hitting capacities for all compact sets, but just for a limited number of compact sets $K^n = \{K_1, \dots, K_n\}$, which we call the structuring elements, since the whole information of the structure of the random field will be captured by the hitting capacities for these sets. The aim of the theory of random closed sets is to estimate the model type and the model parameter such that the observed scene is a typical realization of the RCS.

The likelihood function

Estimates $\hat{T}(K_j)$ of the hitting capacities can be obtained by placing each structuring element in K^n at N (sampling number) positions $\{p_i\}_{i=1}^N$ inside the observation window, each time evaluating the event: K_j hits (or not) the random field Ξ :

$$\hat{T}(K_j) = \frac{k_j}{N}.$$

In order to estimate the model parameter, assuming the model type to be known, we must find the likelihood of observing a given set of hitting capacities, given the model parameter: $p(\hat{T}(K^n) | \theta_i)$. The model parameter θ_i can then be obtained as the maximum likelihood estimate:

$$\hat{\theta}_i : \left. \frac{\partial \ln(p(\hat{T}(K^n) | \theta_i))}{\partial \theta_i} \right|_{\theta_i = \hat{\theta}_i} = 0. \quad (1)$$

If the N events leading to the determination of the sample estimates $\hat{T}(K_j)$ are mutually independent, the probability of a given number of hits k_j (for a given structuring element K_j) follows a binomial distribution:

$$p(k_j|N) = \binom{N}{k_j} T(K_j)^{k_j} (1-T(K_j))^{N-k_j},$$

where $T(K_j)$ is the hitting capacity, dependent on the model parameter. The variance of $p(k_j|N)$ is $\sigma_{k_j}^2 = NT(K_j)(1-T(K_j))$. In order to guarantee that the individual hitting events are mutually independent it is important to choose an appropriate sampling number N and the positions at which the events (hit or not hit) are evaluated. This number depends on the size of the observation window, on the structuring element and on the RCS model. If the samples are taken on a (appropriately chosen) regular grid in the observation window we can for Boolean models identify a lower bound to their optimal number. The worst consequence of small sampling numbers (below the optimum number) is that we do not exploit all the information that we can retrieve via the hitting capacities. We can thus characterize the likelihood of a hitting capacity (or equivalently the number of hits k_j given the model parameter as

$$p(k_j|\theta_i, N) = \binom{N}{k_j} T_{\theta_i}(K_j)^{k_j} (1-T_{\theta_i}(K_j))^{N-k_j}.$$

Under the assumption that the estimates for all hitting capacities are independent we obtain the joint likelihood as

$$p(k^n|\theta_i) = p(k_1|\theta_i) \dots p(k_n|\theta_i),$$

omitting the dependency on the sampling number N . The likelihood function $p(\hat{T}(K^n)|\theta_i)$ is obtained from this density by a simple scaling/normalizing operation:

$$p(T(K^n)|\theta_i) = \prod_j p\left(\frac{k_j}{N}|\theta_i\right).$$

Choice of structuring elements

An important issue concerning the estimation of the model parameters is the choice of the set K^n . The problem is to determine (i) how many structuring elements should be used and (ii) their shape. We assume now that the maximum likelihood estimate given by equation (1) is unbiased. In this case a lower bound of the covariance of the estimate is given by the inverse of the Fisher information matrix:

$$J_{k^n} = -E\left\{\nabla_{\theta} \left(\nabla_{\theta} (\ln(p(k^n|\theta)))\right)^T\right\}$$

We searched for the structuring elements that maximize the determinant of J_{k^n} . For a simple Boolean model, with $\theta=(\lambda, r)$, Poisson point process of intensity λ , and whose objects are

compact discs of radius r , and restricting the analysis to structuring elements that are squares of varying side d , the hitting capacities can be easily shown to be given by

$$T_\theta(K_j) = 1 - \exp(-\lambda(d_j^2 + 4d_j r + \pi r^2))$$

And the Fisher information matrix is obtained from using this expression on the binomial distribution found before, and computing

$$J_{k^n} = -E \left\{ \begin{bmatrix} \frac{\partial^2 \ln(p(k^n|\theta, N))}{\partial^2 \lambda} & \frac{\partial^2 \ln(p(k^n|\theta, N))}{\partial \lambda \partial r} \\ \frac{\partial^2 \ln(p(k^n|\theta, N))}{\partial \lambda \partial r} & \frac{\partial^2 \ln(p(k^n|\theta, N))}{\partial^2 r} \end{bmatrix} \right\} = J_{k_1} + \dots + J_{k_n}$$

It is trivial to show that a single structuring element is not sufficient in order to estimate this RCS model, since it leads to a singular Fisher information matrix.

We assume that the first structuring element k_1 is a single point, yielding

$$\left| J_{k_1, k_2} \right| = c_1 a_1^2 c_2 b_2^2 + c_1 b_1^2 c_2 a_2^2 - 2c_1 a_1 b_1 c_2 a_2 b_2,$$

where

$$c_i = \frac{N \exp(-\lambda a_i)}{1 - \exp(-\lambda a_i)}; a_i = d_i^2 + 4d_i r + \pi r^2; b_i = 4d_i \lambda + 2\pi \lambda r.$$

The other structuring element k_2 is a square of side d_2^* which is the solution of:

$$\left. \frac{\partial |J_{k_1, k_2}|}{\partial d_2} \right|_{d_2 = d_2^*} = 0.$$

It is easy to show that an unique positive solution $d_2^* > 0$ exists, which can be easily found numerically.

We addressed next the following question: What is the information-gain if we add a third structuring element of side d_3 ?

The Fisher information matrix is now:

$$J_{k^3} = J_{k^2} + J_{k_3} = G_{k^2} (\mathbf{I} + G_{k^2}^{-1} J_{k_3} G_{k^2}^{-1}) G_{k^2},$$

where G_{k^2} is a symmetric square-root of J_{k^2} . The determinant of J_{k^3} is

$$\begin{aligned} |J_{k^3}| &= |J_{k^2}| \left| \mathbf{I} + G_{k^2}^{-1} J_{k_3} G_{k^2}^{-1} \right| \\ &= |J_{k^2}| \left(1 + \left\| G_{k^2}^{-1} \sqrt{c_3} \begin{bmatrix} a_3 \\ b_3 \end{bmatrix} \right\|^2 \right). \end{aligned}$$

In the above equation, $G_{k^2}^{-1} J_{k_3} G_{k^2}^{-1}$ is the information gain provided by the third structuring element. Study of this gain as a function of the d_3 indicates that the optimal third structuring

element always coincides with either $K_1=sq(d_1=0)$ or $K_2=sq(d_2^*)$, depending on which is of the two corresponding hitting probabilities is affected by a larger uncertainty, see Figure 2 showing a plot of the gain for a specific Boolean model. This study indicates that for the type of Boolean models two distinct structuring elements provide all required information, that grows with the number of samples that are used to estimate the corresponding hitting capacities. For this reason, in all numerical studies presented in subsequent sections of this paper, we use $n = 2$.

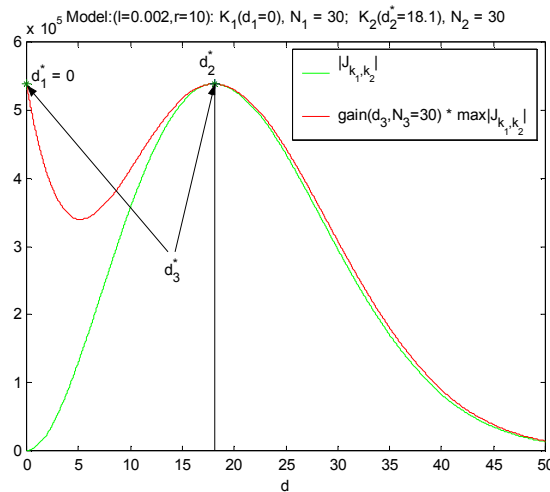


Figure 2: Determinant of Fisher matrix for $n=2$ and the information gain of an additional third shape, showing that its optimal size is either $d_1=0$ or d_2^*

References

- [Rissanen89] Jorma Rissanen, *Stochastic Complexity in Statistical Inquiry*, World Scientific, Series in Computer Science—Vol. 15, 1989.
- [Cover91] Thomas Cover and Joy A. Thomas, *Information Theory*, John Wiley & Sons, Wiley Series in Telecommunications, 1991.
- [Tenas01] Albert Tenas, *Unsupervised Segmentation of Images Based on the Kullback Distance*, Dissertation, Laboratoire d'Informatique, Signaux et Systèmes de Sophia Antipolis, UPC Barcelona, June 2001.
- [Dembo92] Amir Dembo and Ofer Zeitouni, *Large Deviations Techniques and Applications*, Jones and Bartlett Publishers, Inc., 1992.
- [Lloyd82] S. P. Lloyd. Least squares quantization in PCM. *IEEE Trans. Information Theory*, 28: 129-137, 1982.
- [Tenas2001a] Albert Tenas, Maria-João Rendas and Jean-Pierre Folcher, Image Segmentation by Unsupervised Adaptive Clustering in the Distribution Space for AUV Guidance Along Sea-bed Boundaries using Vision, *Proc. Oceans'2001*, November 5-8 2001, Hawaii, USA.
- [Rolfes2001] Stefan Rolfes and Maria-João Rendas, Statistical habitat maps for robot navigation in unstructured environments, *Proc. Oceans'2001*, November 5-8 2001, Hawaii, USA.
- [Stoyan95] D. Stoyan, W.S. Kendall, J. Mecke, *Stochastic Geometry and its Applications*, John Wiley & Sons, 1995.