29th International Conference on Informatics for Environmental Protection (EnviroInfo 2015)
Third International Conference on ICT for Sustainability (ICT4S 2015)

Environmental Monitoring of Continuous Phenomena by Sensor Data Streams: A System Approach Based on Kriging

Peter Lorkowski
Institute for Applied Photogrammetry and Geoinformatics
Jade University of Applied Sciences
Oldenburg 26121
Email: peter.lorkowski@jade-hs.de

Thomas Brinkhoff
Institute for Applied Photogrammetry and Geoinformatics
Jade University of Applied Sciences
Oldenburg 26121
Email: thomas.brinkhoff@jade-hs.de

***Abstract*—Environmental monitoring as technological endeavour has to deal with limitations in respect to transmission capacities, computational resources and storage space. Despite the progress in ICT, those limitations remain significant because of rising demands like real-time monitoring on the one hand and increased amount of observations on the other. The geostatistic method of kriging, originally developed to manage spatial uncertainty, provides elaborate means for spatio-temporal interpolation and is therefore the method of choice wherever continuous phenomena are to be monitored by discrete observations. With its associated confidence estimation, the method can also be exploited to solve other problems like fusion of sub-models for continuous real-time updates or filtering massive sensor streams. In this work, we suggest a system, substantially based on kriging, to filter, process, interpolate, monitor and archive sensor data streams. We incorporate several algorithmic and technical solutions into this system and evaluate them by simulated and real data.**

***Index Terms*—environmental monitoring; sensor data streams; continuous phenomena; geostatistics; kriging**

I. Introduction

Environmental observation data can be used to feed scientific models which explore causations, make predictions and thus support decisions [12]. Very often the environmental data describes continuous phenomena like temperature, radiation or substance concentration. Those phenomena can be modeled as fields that are continuous in space and time. They show different dynamism in each of those dimensions depending on the variable observed. So the models for soil pH level, air pressure, ocean temperature and fine dust pollution show significant differences in their spatio-temporal dynamics.

For a measurement campaign the sampling must be chosen carefully and in accordance with the dynamism characteristics of the observed phenomenon. On the one hand, the field observation will always remain limited and values for non-sampled positions in space and time have to be estimated by interpolation. On the other hand, due to technological progress, the amount of available sensor observations can become overwhelming while at the same time inadequately distributed in space and time. Intelligent mechanisms are necessary to filter out observations of redundant character, especially when the observation platforms are mobile and autonomous (like traffic participants, drifting buoys, etc.).

The field of geostatistics provides elaborate means to deal with observations of spatio-temporally continuous phenomena and associated problems [19], [2], [7]. It can be used to quantify the dynamism of an observed phenomenon, to interpolate at arbitrary positions in space and time, but also to estimate the confidence of each interpolation. It covers the stochastic part of a continuous phenomenon, but can also help to detect systematic or deterministic influences [19]. Those can subsidiarily be considered, e.g. by applying fluid dynamics.

The overall goal of such modeling is an appropriate representation of the observed phenomenon for scientific purposes like climate research, but also for commercial purposes like fishing or agriculture. When considering temporal dynamism, such representation can be described as a scenario or movie of continuous value distribution (e.g. as known from meteorological animations of atmospheric pressure). This representation can be computed from discrete observations at arbitrary spatial and temporal resolution, e.g. by means of the geostatistical method of kriging. Diverse extracts like snapshot maps, time series at fixed positions or arbitrary spatio-temporal aggregations can easily be derived. Besides the interpolated value itself, kriging also provides its estimation variance, which can be indispensable for applications where the confidence range of an information is crucial.

In the context described above, we propose a system architecture to process sensor data streams describing continuous phenomena. We exploit the capabilities of kriging to target diverse problems of environmental monitoring like inhomogeneous observation distributions, massive data feeds, (near) real-time requirements, error assessment and efficient archiving and retrieval. Apart from the general system design we provide experimental results concerning the introduced methods that have been applied on synthetic models as well as on real measurement data.

The remainder of the article is structured as follows: In the next section we introduce the problem domain of monitoring continuous phenomena. Works related to this field are dis-

© 2015. The authors - Published by Atlantis Press

cussed in III. The major features of the geostatistical method of kriging are revisited in IV. Based on this review, we introduce our system architecture in V and present experimental results about some of its components in VI. In the light of those results, we discuss the applicability and limits of our approach in VII before we finally conclude our work and prospect future effort.

II. The Problem Domain: Monitoring of Continuous Phenomena

A. Characteristics of Continuous Phenomena

Many aspects of our environment have a continuous character. Whether they are of physical, chemical, biological or even social kind [7, p. 11], we often find continuous variability over space and time instead of abrubt changes, which might indicate some kind of anomalies like faults or fractures (e.g. in geology). Mostly, we find a varying value caused by some source of force like warmth by sun radiation, soil nitrate content by overfertilization, oxygen deficiency by algae groth or air pollution by commuter traffic.

The distribution of such a variable in space and time is often the result of highly complex or even chaotic processes. Because it is impossible to describe such processes completely deterministically, the resulting variable can be considered as (at least partly) random [11, p. 198ff.], [19, p. 47]. So, depending on the phenomenon observed, prediction can be performed using deterministic, statistic or mixed approaches. For instance, while generally modeling a phenomenon as random process by kriging, deterministic portions like trends, anisotropies or periodicities can be incorporated in the calculation [16], [19].

B. Discrete Observations and Continuous Representation: Geostatistical Modeling

The observation of continuous environmental phenomena will always be limited. In the best case, it is sufficient to support a hypothesis or make reliable predictions about non-sampled positions or future states. The observation design can only be a pragmatic compromise between accuracy and cost. To capture the scales of variation of the particular phenomenon, some test sampling at different distance stages might be necessary [19]. If appropriately chosen, those initial observations reveal the characteristics of the phenomenon by their variogram (see IV-B).

When applying an interpolation algorithm on discrete observations, in most cases the desired result is a georeferenced and regular raster or grid representing the interpolated variable by pixel values. When displayed as color (e.g. greyscales), the distribution of values can intuitively be interpreted (see Figure 1).

When temporal dynamism is considered, an animated representation or movie (like in rain radar forecasts) is appropriate [8], [21]. Abstractly speaking, such a movie or scenario provides a seamless coverage of a variable at a particular spatial (pixel size) and temporal (frame rate) resolution. Principally, the real existing physical field of interest can also be seen as

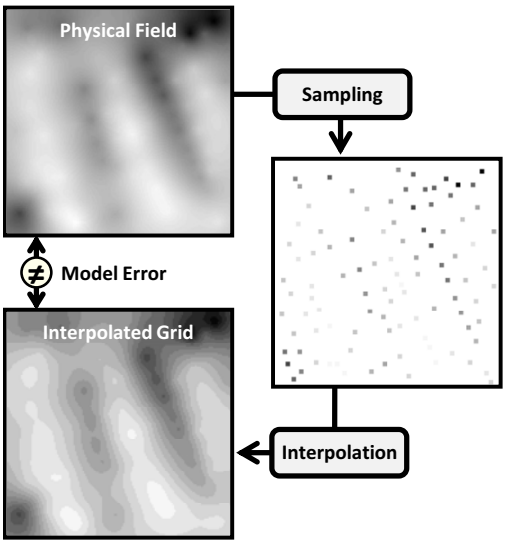

Fig. 1. Monitoring continuous phenomena

such a grid, but as one of infinitesimal resolution in space and time. It remains unknown and can only be sampled by finite discrete observations. Thus, the principal objective of environmental monitoring might be described as follows: “Find an appropriate combination of sampling configuration and interpolation algorithm so that the resulting grid approximates the real world physical field accurately enough to fulfill the given task.” Depending on the task, sparse sampling with simple interpolation might suffice, or dense observations of diverse variables and highly complex calculation models might be necessary. In any case, there will always remain some discrepancy or model error between the real phenomenon and the interpolated grid (see Fig. 1).

This discrepancy, which can only be revealed exactly when working with synthetic phenomena, should in any case be estimated and checked against the monitoring objectives.

Since continuous environmental phenomena can, at least partly, be seen as random (II-A), in the process of monitoring such phenomena (Figure 1) geostatistical principles and rules have to be considered carefully.

III. Related Work

Generally speaking, environmental monitoring of continuous phenomena deals with many issues, of which some are:

- deterministic physical models,
- sensor models (A/D conversion),
- signal processing,
- sensor communication / sensor web enablement (SWE),
- data processing / data analysis, and
- visualization.

Diverse approaches have been suggested to face the requirements in this area, especially in the field of data processing and data analyis. In the following, we briefly introduce some of the work we found relevant and set our approach in relation:

Whittier et al. [21] suggest a new design of a Data Stream Engine (DSE) that is based on k Nearest Neighbors (kNN)

and spatio-temporal inverse distance weighting (IDW). It uses main-memory indexing techniques to address the problem of real-time monitoring of massive sensor measurements. In contrast to this approach, in our proposed system we avoid fixed sized sub-models based on temporal intervals, but aim at a flexible segmentation that can adapt to a streamed data rate. By merging sub-models continuously, we also consider old observations as long as no better information is available. This might be especially important when observations are inhomogeneously distributed in space and time.

Appice et al. [1] inspect trend clusters in data streams and discuss techniques to summarize, interpolate and survey environmental sensor data. Since one main application is the detection of outliers within a rather low dynamic phenomenon (solar radiation), the approach allows a coarse approximation by clusters of similar values. For our purpose, a rather smooth representation of each state is desirable.

Walkowski [18] uses the kriging variance to estimate a future information deficit. In a simulated chemical disaster scenario, mobile geosensors are placed in a way that optimizes the prediction of the pollutant distribution. Instead of optimizing the observation procedure itself, we exploit the kriging variance to fuse sub-models in order to optimize calculation and to filter out redundant observations.

Katzfuss and Cressie [13] decompose a spatial process into a large-scale trend and a small-scale variation to cope with about a million of observations. This solution is an option for optimizing very large models, but is not helpful for our sequential approach with its real-time specific demands.

Osborne et al. [16] introduce a complex model of a gaussian process regression (synonym for kriging) that incorporates many factors like periodicity, measurement noise, delays and even sensor failures. Similar to our work, sequential updates and the exploitation of previous calculations are performed, but here on a matrix algebra basis. It uses kriging with complex covariance functions to model periodicity, delay, noise and drifts, but does not consider moving sensors.

Wei et al. [20] use a k-d tree-based method to partition big datasets into child data groups which can be processed in parallel by kriging. This method is particularly suitable to cope with big datasets, but does not consider the temporal dimension. In contrast to this approach, we divide big datasets not by the spatial, but by the temporal dimension to reflect the dynamism of the phenomena. By continuously fusing subsequent partial models, we follow a principle of seamless actualization and gradual decay of outdated information.

In the context of Data Stream Management [9], the main objectives of monitoring are the optimization of transmission, aggregation and pattern recognition of sensor data. While providing recursive methods for simple statistical parameters like mean or deviation[10, p. 30ff], the handling of the far more complex geostatistical properties is not covered here. Our contribution in this context is a method for successive and smooth update of continuous models while providing the structure for complex event detections (exceeded threshold considering variance) in a continuous phenomena context.

IV. Geostatistics: Kriging

In a spatial context, the method of kriging has developed "to be synonymous with 'optimal prediction"' [6]. This method is typically used to interpolate between discrete samples (e.g. of ore concentration) on a two-dimensional plane, but it can also be used for transects, time series, 3-dimensional spaces or spatio-temporal reference systems (with 1, 2 or 3 spatial dimensions) [8].

Kriging reflects the fact that observations proximate in space and time tend to be more similar than distant ones; a phenomenon also known as autocorrelation. A variable of this character is called regional [7], [8] and can describe dynamic continuous phenomena. Since many environmental phenomena are of this kind, kriging is an important methodology here.

A. General Principle

Kriging can be seen as "a transposition of multiple regression in a spatial context." [17, p. 15]. It is a tool to model phenomena or associated data that partially are the result of a random process [7, p. 29]. It exploits the fact that the general distribution of a regional variable in dependency of spatial proximity can be described prior to the actual statistical calculation. This Bayesian approach, in contrast to classical frequentist inference, incorporates prior knowledge about the process into the model and thus tries to achieve the best possible prediction [8, p. 27ff]. This knowledge, in the case of kriging, is expressed through the variogram.

B. Variogram and Covariance Function

Within geostatistics, the empirical variogram, the theoretical variogram and the covariance function represent the key elements. Those elements have been generated for a small sample data set and are presented in Figure 2.

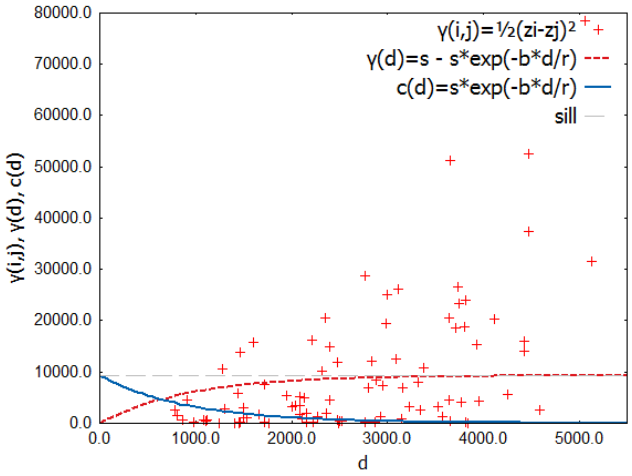

Fig. 2. Empirical variogram, theoretical variogram and covariance function

The empirical (semi-)[1]variogram depicts the actual statistical properties of a sample by plotting the halved squared value

[1]It is the halving of the differences where the term "semi" comes from; we will use the short form "variogram" here.

differences of a pair of observations against their distance. The theoretical variogram generalizes the distribution of those points as mathematical function. The covariance function can be derived from the theoretical variogram as its mirror image about a line parallel to the abscissa. It expresses the pairwise (auto-)correlation between observations and is the function actually required for the kriging calculation. For appropriate modeling, the theoretical variogram (and its associated covariance function) needs to be adjusted to the empirical variogram. A common approach is to create the empirical variogram from the sample set, from the many theoretical variogram models choose the one that best represents its statistical characteristics [19] and then fit its parameters to the actual data set (e.g. by least squares matching, [5]).

The general task of the theoretical variogram and its associated covariance function is to express the degree of correlation between two observations. In the simplest case, the distance between two points in Euclidean space determines the correlation of the observed variable, which is obvious for continuous phenomena: temperature or air pressure tends to be similar for nearby positions. When considering anisotropy, not only the distance but also the direction of the vector connecting those points determines their correlation. This might reflect the distribution of air pollutants at constant wind. Periodic patterns in space or time can be expressed by a fluctuating variogram and thus be considered in the statistical processing. As long as it can be expressed mathematically, any other correlating aspect can be incorporated into a covariance function of arbitrary complexity, as long as its definiteness is fulfilled [19]. On the other hand, care should be taken not to model properties that do not exist in reality (over-parameterisation [19, p. 95]).

While originating from rather static applications (like ore-reserve estimation), the method of kriging is increasingly used to model temporally dynamic phenomena ([8], [1], [22]). Since the dynamism of a phenomenon usually differs in space and time (see I), those dimensions should be handled differently in the corresponding covariance function. As can be seen in Figure 3, the correlation is at its maximum when spatial and temporal distance between two positions is zero and decreases when increasing either value. When adjusted appropriately, the shape of the functional surface reflects the statistic characteristics of the observed phenomenon. Specifically, it relates spatial and temporal variation to each other [8].

C. Interpolation Values and their Variances

One of the central advantages of kriging over other interpolation methods is that it provides an estimation of uncertainty for each interpolated value by the kriging variance. In many cases, in addition to the interpolated value, its error assessment is of interest. For example, when the observed pollution of a district is just slightly below a critical threshold, additional observations could be necessary when confidence is low. Just as the value itself, the average conficence for a district can be derived from the corresponding variance map (see Fig. 4).

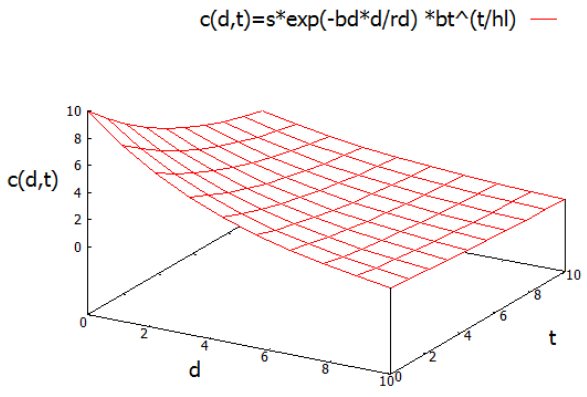

Fig. 3. Spatio-temporal covariance function

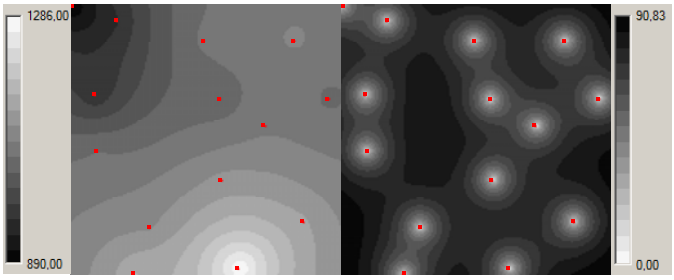

Fig. 4. Kriging results: value map and variance map

In the context of our system, we consider the variance as coherent element of an interpolation result and exploit its explanatory power for diverse purposes. We introduce the fusion of sub-models and the filtering of input data streams here (see V-B and V-C).

V. System Architecture

A general system design for monitoring continuous environmental phenomena with associated functions and services will be sketched in the next subsections. It aims at an overall solution with (near) real-time processing and archiving requirements by exploiting the capabilities of kriging.

A. Overview

As general guidance for the suggested architecture of a monitoring system (see Fig. 5), we presume a continuous stream of georeferenced and timestamped observations of one or more environmental variables. This data stream might be provided according to the specifications defined for the sensor web enablement (SWE) [4]. We further presume some basic requirements to be fulfilled by the system like (near) real-time monitoring by web mapping, filtering of the input stream in case of high data volume, recognizing and reporting predefined critical states and storing a compressed extract of the scenario to an archive.

For many features of the data stream engine sketched here, kriging provides useful capabilities. So for a continuous

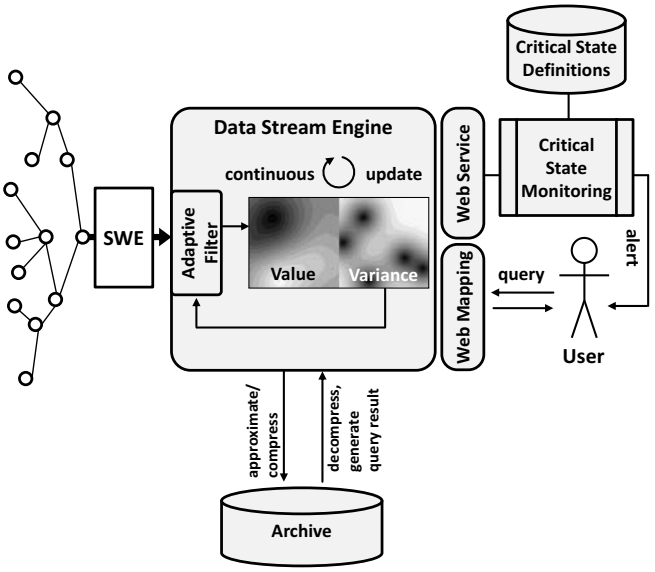

Fig. 5. System Architecture

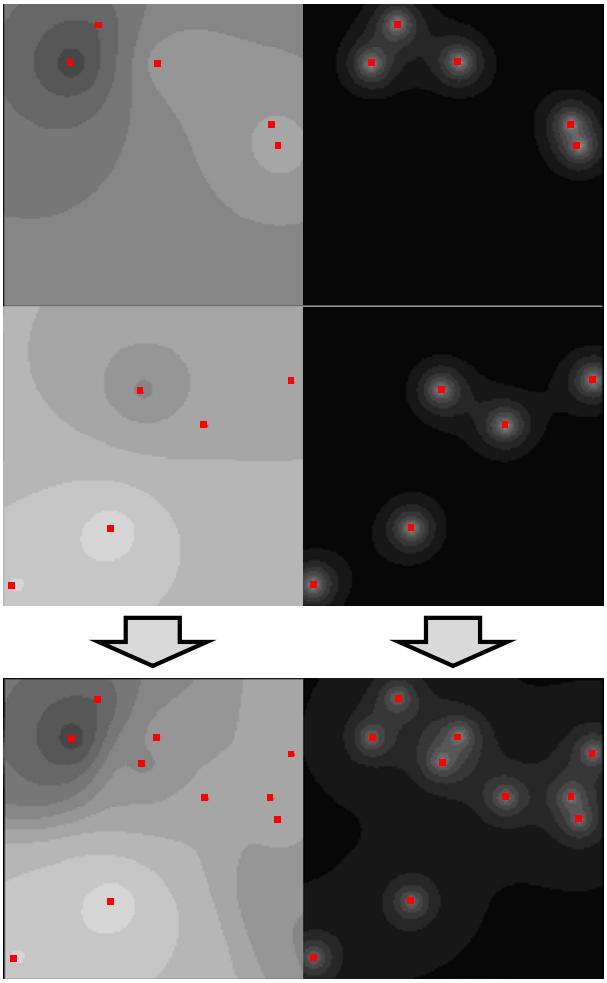

Fig. 6. Fusion of sub-models based on kriging variance

update of a current map, new incoming observations should be integrated seamlessly into the previous model. For that, the new sub-model is merged with the former sub-model by performing a fusion of both models based on their corresponding variance maps. Thus, the new model smoothly and seamlessly overwrites the old model according to its regional variances (see V-B).

In a real-time processing environment, the data volume might overtax the computational capacities of such system. In this case, the input data stream must be thinned out without fundamental information loss. The subsequently updated variance map can express the significance of new observations by indicating if it is situated in a region yet well or poorly observed. Thus, observations that are most likely redundant can be filtered out and do not burden the costly kriging process (see V-C).

The definition of critical states would typically specify a region for which a particular variable should not exceed a threshold. Another aspect might be insufficient observation indicated by high variance map values for that region. Values narrowly below the threshold combined with high variance point to a high exposure risk (see V-D).

In view of continuously increasing amounts of sensor data, an appropriate compression strategy should be an integral part of any monitoring system (see V-E).

B. Fusion of Sub-Models

For a real-time monitoring system, new observations should be integrated seamlessly into the continuously updated grid or map. Since the sampling might be dispersed in space and time, previous observations should not be ignored in the new map, which would be the case with fixed time intervals. On the other hand, considering all observations - previous and new ones - for each new kriging calculation would inevitably exceed the computational capacities, since calculation effort grows cubically with the number of samples [1, p. 63]. So, while sticking to the unquestionable advantages of kriging (interpolation quality, estimation of uncertainty), a strategy to cope with its computational complexity for massive data is necessary [15].

As can be seen in Figure 6, the main idea of our approach is to use the inversed variances as weights when fusing two grids generated from different sub-sets of samples of a region.

When applied sequentially, this method tendentially "overwrites" the former grid only in regions where the new grid´s variance is significantly lower. The variance maps themselves are also fused (basically by addition, taking into account temporal decay), thus representing the confidence distribution of the new model and determining its weighting for the next fusion step.

The fusion process is performed for each grid cell by deriving the weight p from its standard deviation with

$$p = \frac{1}{var^d}, \tag{1}$$

where var is the variance of each grid cell calue derived from kriging and d is the exponent controling the decay of

weight due to that variance. This factor can be adjusted according to the spatio-temporal disperison of the given dataset. With values and weights for each grid cell, the merged model values $x_{[n+1]}$ can be derived from the current sub-model values $x_{[n]}$ and previous model values $x_{[n-1]}$ by

$$x_{[n+1]} = \frac{x_{[n]} \cdot p_{[n]} \ + \ x_{[n-1]} \cdot p_{[n-1]}}{p_{[n]} + p_{[n-1]}}. \quad (2)$$

If the temporal dimension is considered in the model, the weight of the previous model $p_{[n-1]}$incorporates the temporal decay factor

$$tdec = bt^{\frac{t_{[n]} - t_{[n-1]}}{rt}}, \quad (3)$$

where bt represents the fraction that the original has decayed to after rt time units and $t_{[n]} - t_{[n-1]}$ is the time lag between the reference timestamps of the sub-models. Wheras the deviation within each sub-model is derived as linear combination of spatio-temporal covariances towards its observations, the temporal decay between the models as a whole is expressed by this factor.

Finally, we generate the variance map for the fused model by

$$var_{[n+1]} = \frac{1}{\frac{1}{var_{[n]}} + \frac{1}{var_{[n-1]}}}, \quad (4)$$

where the variances of the previous model $var_{[n-1]}$ might also incorporate a temporal decay factor similar to (3).

This methodology considers sub-models as statistically self-describing information entities and in principle allows the fusion of any combination of them, regardless their spatial or temporal distribution. This is a powerful principle for processing, archiving and analyzing tasks in a monitoring environment.

C. Data Stream Filter

When using observations of autonomous sensor platforms (e.g. drifting buoys, private transport vehicles), a spatial (and/or temporal) concentration of observations can easily occur so that regions might be observed redundantly. Including all those clustered observations will encumber the model calculation without significantly improving its accuracy. In a monitoring environment dealing with massive data streams, an appropriate filtering can become indispensable to ensure real-time functionality.

With the method of kriging, an estimation of variance (or confidence, respectively) is given for each interpolated point in the result grid (see Fig. 4). Being derived from a linear combination of the (spatio-temporal) covariance function applied to the current point, it reflects the quality of determination according to the spatial and temporal proximity to observations done so far. In regions of high confidence (or low variance), it is unlikely that new observations differ substantially from the current model, except for outliers or anomalies. Thus, the variance map of the current model can be used to assign a priority ranking among new observations. This ranking can then be used to dynamically adapt to changing data rates, while at the same time striving for maximum overall confidence.

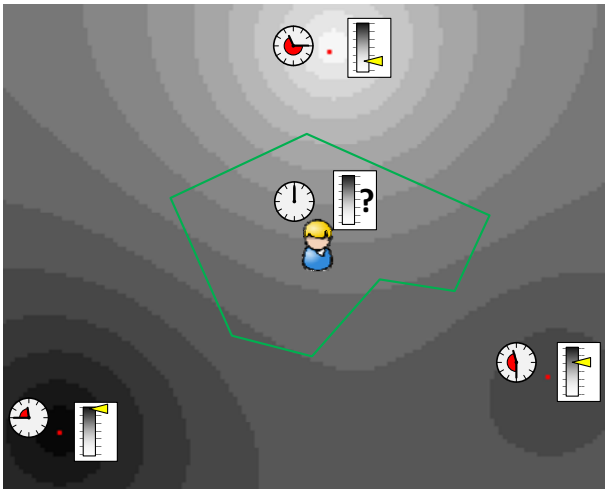

Fig. 7. Spatio-temporal interpolation and aggregation

Since the decay of correlation caused by increasing spatial and temporal distance is usually different for each data set, the filter should adapt to those circumstances. As will be shown in VI-C, the statistical characteristics of a data set can be revealed by its empirical variogram, which can be used to adjust the parameters of the covariance function. Since the covariance function determines the creation of the variance grid, the behavior of an associated filter is, in general, also determined by it.

D. Complex State Definitions

When used as a real-time monitoring system that generates alerts when thresholds are exceeded, the data stream engine continuously needs to check the current distribution of a regional variable against predefined critical states. Such a state could be defined by a threshold for the estimated value at a particular point. When defined for regions or districts, the average value of the enclosed grid cells could be an appropriate aggregate. When considering temporal dynamism, the daily average or peak value can be used. Any other spatio-temporal aggregation can easily be derived from the grid.

Apart from interpolation and aggregation of the pure value, its estimation confidence might also be of importance in some scenarios. For example, when monitoring a particular phenomenon by autonomous sensors like drifting buoys or traffic vehicles, some important region might be under-observed due to the uncontrollable distribution of the sensors. This situation can be detected by an aggregation of not only values but a combination of values and their associated kriging variances. A supplementation of the current model by additional observations in that region could be induced. An optimized distribution of controllable sensors can be performed this way [18].

The general idea of complex state definitions is depicted in Figure 7. It illustrates a fundamental problem of environmental

monitoring where the spatial and temporal distribution of observations do not cover the information requirements. Appropriate interpolation techniques need to be applied, typically on a grid of sufficient resolution. Aggregations of interpolated values (and variances) for predefined regions can easily be derived from intersecting raster cells (Fig. 7).

More generally speaking, kriging combined with complex state definition can help to bridge the spatio-temporal gap between available observations and required knowledge. This will be a feature of growing importance for coming monitoring systems dealing with massive sensor data.

E. Compressed Archiving

Besides the real-time capabilities mentioned above, an efficient archiving and retrieval of the aquired monitoring data is desirable. Especially for scientific purposes where correlations between the observed phenomenon and other data (e.g. weather, traffic, etc.) are of interest, a comfortable access to archived monitoring data is crucial.

Within our system, we store the data in self-descriptive units or sets of observations given for a confined spatio-temporal area. The interpolation of values for arbitrary positions in space and time within this area is achieved through kriging. Thus, those units can be seen as movies about the spatio-temporal distribution of values and their variances and can be interpreted intuitively (see IV-C). Nevertheless, for long term archiving we consider the movie format (raster sequences with particular resolution and frame rate) as unappropriate for the following reasons:

- For a given campaign the generated “movie scene” will in most cases afford considerably more storage space than the set of observations it is based on, even if compressing techniques are applied.
- The movie scene does not contain the original, but only derived data. If an offset error of a sensor becomes known after processing, it can hardly be considered or corrected here.
- The resolution (both: spatial and temporal) is fixed and resampling might be less accurate and more awkward than generating a new model with a different resolution from the original observations.

On the other hand, there is the drawback of the higher interpretation effort when historic observations have to be interpolated by kriging for visualization or aggregation. To cope with this problem, for scenes which are frequented often, the derived movie could be cached.

Alternatively to the original observations themselves (or a deliberately selected subset of them), a new sampling can be performed on the derived model. This resampling should seek for an appropriate compromise between the number of resampling points and the fidelity of the derived model compared to the one derived from the original observations.

Whether original observations or resampling points, the structure of such a point cloud has potential for compression. The basic idea here is to define n-dimensional binary paths for each point within the area observed and thus reduce the necessary memory per point. The compression ratio will depend on the size of the area (given by n-dimensional bounding box) and required numeric precision (e.g. given by number of fractional digits).

On the one hand, the compression technique helps to reduce storage space in archiving scenarios. Additionally, when properly applied, it could also be used to reduce network communication traffic, which is particularly important in wireless sensor networks where transmission of data is usually costly [3].

Besides the compression, the introduced technique supports progressive decoding of stored data. Because not all applications need the full accuracy of a data set, it can be loaded progressively until the required accuracy is achieved. This is possible because of the binary tree structure for each dimension. With every level, the precision in each dimension is doubled, until it is sufficient for the particular task. Thus, a significant reduction of network traffic and computation effort can be achieved.

VI. Experimental Studies

In this section, we present some (first) investigations of selected concepts, methods and algorithmic solutions within the presented framework in order to demonstrate its functionality and benefits.

A. Synthetic Data Generation

An objective judgement of the introduced methods for interpolation, filtering, fusion and compression is rather limited when using real sensor data. The distribution of a particular variable is principally unknown in unobserved areas and therefore the methods associated with the interpolation of those areas can hardly be evaluated. Hints for the overall interpolation quality can be found by cross-validation [8, p. 39ff], but such evaluation will strongly depend on the density and consistency of the given observational data.

Alternatively, a completely gapless synthetic reference model of a dynamic continuous phenomenon can be created. Sampling of arbitrary spatial and temporal density can be performed thereon. From such a synthetic sampling set, a grid model, ideally of same spatio-temporal resolution as the reference, can be interpolated. A quantification of the discrepancy between the reference and the derived model is straightforward (see Fig. 1). Such discrepancies can also be quantified for a whole region by the root mean square error (RMSE). Moreover, visualizing those discrepancies as a grid might reveal weaknesses of the procedure used and give hints for improvement, which can immediately be tested and exactly evaluated on the given reference.

Simulated fields of random processes can be created by using covariance function definitions as filters on grids of pure white noise [19]. For each pixel position, its new filtered value is calculated as weighted mean of the surrounding pixels, whereas the selected covariance function determines the weight for the corresponding distance. This process smoothes out the pure noise and, when applied on 2 dimensions,

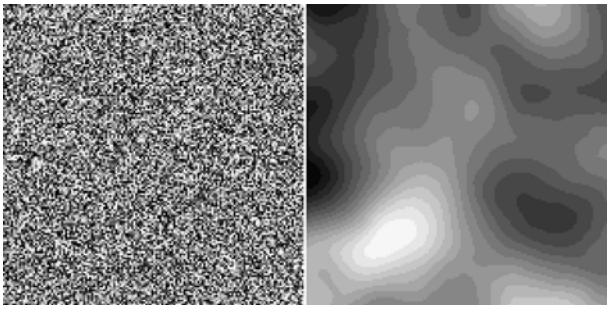

Fig. 8. Random map (150 x 150 pixels) (r) generated by applying a covariance function (spherical, range: 45 pixels) as filter on a pure white noise image (l).

creates a relief that expresses the statistical properties of the covariance function used (see Figure 8).

The similarity of the resulting synthetic model with natural continuous phenomena is striking and reflects the appropriateness of the method of kriging for such structures. In principle, the method is used reversely here. By applying different covariance functions, also different models can be created, representing the function properties through their structure. A temporal dimension can easily be added by creating a movie (II-B) of white noise (e.g. with one image per second) and applying a spatio-temporal covariance function as filter.

This method provides unlimited possibilities for the creation of n-dimensional scenarios on which observations (regularly or randomly dispersed) can be performed. These can be interpolated and compared to the reference by RMSE. So the process of sampling and interpolating, fusing and filtering can easily be evaluated and optimized.

B. Fusion of Sub-Models

The fusion of sub-models (see V-B) was designed to reduce the computational complexity of kriging and to allow a seamless integration of new sets of data in a real-time monitoring environment [15]. To test its operability, we applied the method on a synthetic reference grid model. Therefore, we randomly spread 100, 200, 300 and 400 points over it. From those observations we interpolated a result grid of the same extension and resolution as the reference model. This was done

1) using all observations in one kriging calculation and
2) separating the set into sub-models of 10 observations each, calculating and fusing them sequentially.

As can be seen in Figure 9, the sequenced fusion algorithm provides coarse results immediately and in trend improves accuracy while progressing. The accuracy is expressed as RMSE (Root Mean Square Error) towards the reference model. The master model reaches the highest accuracy in all experiments, but also consumes the most computing time compared to the sequenced method. This nonlinear effect becomes obvious with increasing model size and reflects the computation complexity of $O(n^3)$ [1, p. 63].

C. Drifting Buoy Data

Within oceanography, data about temperature, salinity and current are crucial to create a dynamic model of the ocean.

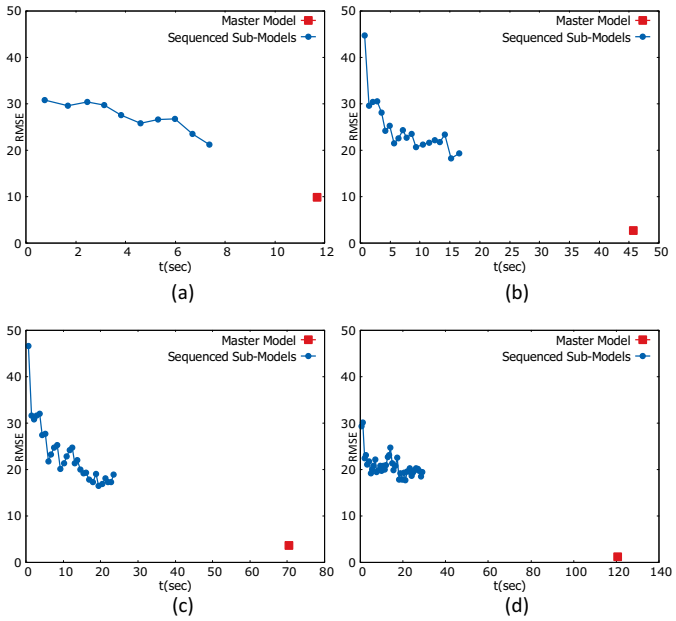

Fig. 9. Experimental evaluation of subsequent fusion algorithm with 100 (a), 200 (b), 300 (c) and 400 (d) samples.

Drifting buoys or floats as used for the Argo program[2] provide such data with help of satellite communication. The data is partly preprocessed and can be obtained as CSV files from Canadian governmental service[3].

According to a first assessment, we consider these data as very appropriate for future application of our system for following reasons:

- Several dynamic continuous phenomena are observed: sea surface temperature and salinity (directly), ocean current (indirectly by positions and timestamps).
- The observations are inhomogeneously distributed in space and time, therefore the testing and evaluation of interpolation and filter algorithms is reasonable.

In a first study, we imported and processed a confined area with our software to assess the characteristics of the data (see Figure 10).

As can be seen from Figure 10, the 150 observations are inhomogeneously distributed, many of them concentrated spatio-temporally due to the observation frequency, thus representing short transects.

For further analysis, we produced the empirical spatio-temporal variogram of the buoy data which is depicted as scatter plot in Figure 11.

Following statements can be derived from the plot:

- Where the axes for spatial and temporal distances meet in their origin, the (semi-)variance of the observation pairs is nearly zero and tendentially grows when increasing the value on either axes.

[2]http://www.argo.ucsd.edu

[3]http://www.meds-sdmm.dfo-mpo.gc.ca/isdm-gdsi/drib-bder/svp-vcs/index-eng.asp

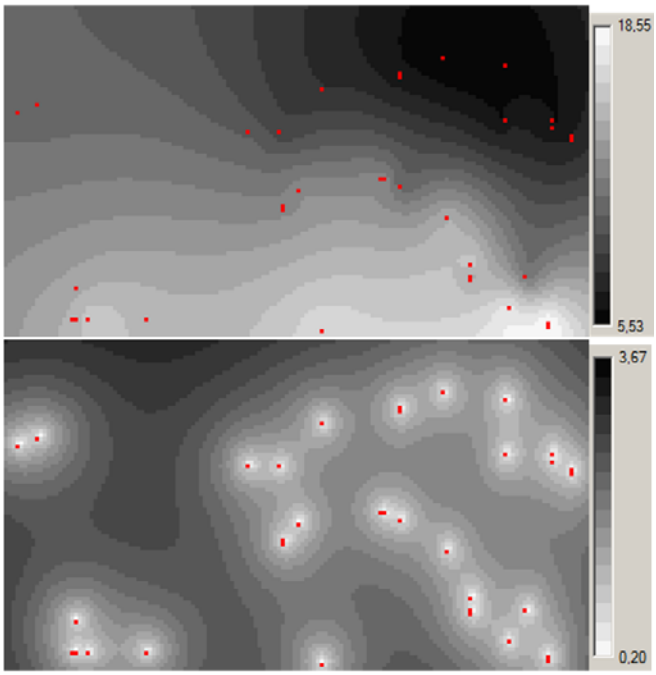

Fig. 10. Kriging of drifting buoy data; interpolated values of sea surface temperature from 150 observations, °C (top) and corresponding deviations (bottom); observed area: north atlantic, N40°-60° W10°-50°, 2011/01/01 00:00h - 10:00h

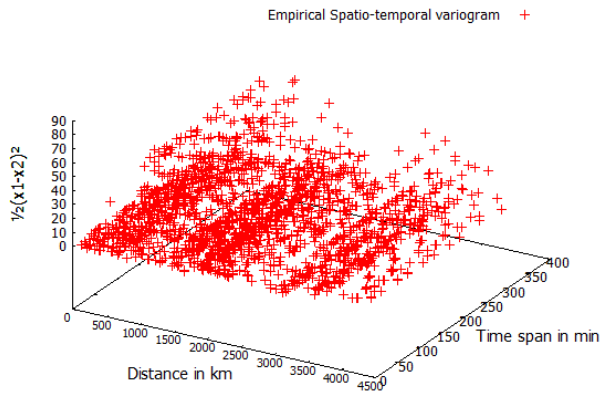

Fig. 11. Empirical spatio-temporal variogram of Argo buoy data

- For the given spatio-temporal area, spatial autocorrelation (see IV) is much more significant than temporal autocorrelation. This is due to the selection of a spatially large (˜4.000 km) but temporally small (10 h) area and would reveal the opposite result if this relation was reversed. With help of this relation, an appropriate adjustment of a data stream filter (see V-C) can be achieved.
- Weak but recognizable linear patterns parallel to the time span axis indicate a dense measurement rate along the buoys trajectories. This effect is caused by a relatively low variation (dependent on the scale of the region) of position and measured value within such a trajectory (or transect).

For spatio-temporal kriging, an appropriate theoretical variogram has to be aligned towards the given empirical variogram by parameter adjustment (IV-B). Since the associated covariance function (see Figure 3) is the mirror image of this theoretical variogram, the trend surface of Figure 11 has to be imagined mirrored at a plane that the (semi-)variance axis is the normal of.

VII. Discussion

Based on an overall design sketch for an environmental monitoring system, we evaluated some of the key components to examine its general feasibility.

In VI-A, we briefly introduced the method of creating synthetic models of random spatial processes based on covariance functions. The general approach is very powerful and can generate models of predefined statistic characteristics. Therefore, it is best suited to test and evaluate a monitoring system under different circumstances.

The experimental fusion of sub-models in VI-B has shown the general feasibility of this approach to cope with massive observational data streams. In contrast to other approaches like spatial segmentation [20] or fixed sized windows [21], it follows the principle of continuity by applying a gradual or fuzzy update method in a spatio-temporal context. Applications like real-time monitoring or dynamic animations can thus be generated easily. A final evaluation of the presented method is not appropriate yet since the merge algorithm is still to be adjusted for different settings.

A first assessment of the applicability of the system to process buoy data from the Argo program in VI-C has already revealed some important insights. The plotting of an empirical spatio-temporal variogram, in addition to the interpolation maps and their associated variances, strongly supports an interpretation and critical evaluation of the given data. The inhomogeneously distributed and, on a global scale, numerous observations are challenging for systems with real-time monitoring requirements. We suppose that our filter mechanism will be helpful here because it adapts to the data dispersion by the associated covariance function.

VIII. Conclusions and Future Work

The monitoring of continuous environmental phenomena has to deal with limitations on several levels:

- Limited density (spatial and temporal) of observations to record a phenomenon.
- Limited knowledge about the (physical) processes that cause the phenomenon.
- Limited resources to transmit, process and store observational data about the phenomenon.

In this article, we suggested methodologies – to a great extent based on geostatistical principles – to counter those limitations. We introduced a system framework to process streaming sensor data. Methods to interpolate, filter, fuse and compress observational data have been presented, mainly by exploiting

the capabilities of kriging. Experimental investigations have shown the feasibility of some core components of the system.

For further evaluation, adjustment and advancement of the system, extensive studies will have to be performed on both, synthetic and real data. While synthetic reference models, as introduced in VI-A, allow otherwise impossible exact comparisons between the real and the interpolated grid model, the work with real world data rigorously reveals any deficiency of the whole environmental monitoring methodology (e.g. distribution and sampling rate, sensor or communication faults, etc.). When appropriately applied, both approaches will complement and stimulate each other.

In the view of the processed Argo buoy data as representative for the observation of continuous phenomena, several hints for further application, progression and extension of the system have already emerged:

- From the empirical spatio-temporal variogram, a theoretical variogram, or covariance function, respectively, has to be derived. Several variogram fitting methodologies as in [5] will have to be applied therefore. An automated and thus objectified procedure as suggested in [19] is desirable here.
- After the general description of the statistic behavior has been deduced by the fitting of an appropriate covariance function, this function also provides the main parameters to adjust the variance based filter. Its precision and efficiency in the detection of redundant observations can be evaluated at different settings when comparing the interpolation effort with the achieved accuracy for filtered and unfiltered data sets.
- Based on the concept of separate calculation of submodels and their subsequent fusion, a dynamic scenario simulation (or movie) of temperature or salinity distribution within the oceans can be generated. The ratio between the spatial and temporal influence expressed by the covariance function will be the key parameter to consider here.
- Derived from the continuous scenario simulation, time series at fixed positions or snapshots of particular regions at particular times can be extracted. From those, aggregations (peak values, mean values, histograms, dynamism indicators) over space and time can be generated. When applied to a whole scene, such aggregations can become part of its metadata and be used for database archiving and retrieval. When updated consecutively on a continuous data stream, a complex alert service can be realized [9].
- The proposed algorithm for compression can be applied for the given drifting buoy observational data. Since it is multidimensional and partially rather coarse (positions in decimal degrees with 3 fractional digits), we expect high compression rates compared to floating number representation. An appropriate representation of integer (buoy IDs) and Boolean (drogue on/off) data types will also be necessary here.

REFERENCES

[1] A. Appice, A. Ciampi, F. Fumarola, D. Malerba. Data Mining Techniques in Sensor Networks: Summarization, Interpolation and Surveillance. London Heidelberg New York Dordrecht: Springer, 2014.

[2] M. Armstrong, Basic Linear Geostatistics. Berlin Heidelberg: Springer, 1998.

[3] J. Barros, Sensor Networks: An Overview. In: J. Gama, M. M. Gaber (Eds.), Learning from Data Streams: Processing Techniques in Sensor Networks. Berlin Heidelberg: Springer, 2007.

[4] M. Botts, G. Percivall, C. Reed, J. Davidson, OGC® Sensor Web Enablement: Overview And High Level Architecture. Open Geospatial Consortium, 2007. Online Document: portal.opengeospatial.org/files/?artifact_ id=25562

[5] N. A. C. Cressie, Fitting Variogram Models by Weighted Least Squares. Mathematical Geology, Vol. 17, No. 5, 1985.

[6] N. A. C. Cressie, The Origins of Kriging. Mathematical Geology, Vol. 22, No. 3, 1990.

[7] N. A. C. Cressie, Statistics for Spatial Data. New York, Chichester, Toronto, Brisbane, Singapore: John Wiley & Sons, 1993.

[8] N. A. C. Cressie, C. K. Wikle, Statistics for Spatio-Temporal Data. Hoboken, New Jersey: John Wiley & Sons, 2011.

[9] J. Gama, M. M. Gaber (Eds.), Learning from Data Streams: Processing Techniques in Sensor Networks. Berlin Heidelberg: Springer, 2007.

[10] J. Gama, P. P. Rodrigues, Data Stream Processing. In: J. Gama, M. M. Gaber (Eds.), Learning from Data Streams: Processing Techniques in Sensor Networks. Berlin Heidelberg New York: Springer, 2007.

[11] E. H. Isaaks, R. M. Srivastava, An Introduction to Applied Geostatistics. New York, Oxford: Oxford University Press, 1989.

[12] E. T. Jaynes, Probability Theory: The Logic of Science. Cambridge, United Kingdom: Cambridge University Press, 2003.

[13] M. Katzfuss, N. A. C. Cressie, Tutorial on Fixed Rank Kriging (FRK) of CO_2 data; Technical Report No. 858. Department of Statistics, The Ohio State University, 2011.

[14] M. A. Osborne, S. J. Roberts, A. Rogers, I. R. Jennings, Real-Time Information Processing of Environmental Sensor Network Data Using Bayesian Gaussian Processes. ACM Transactions on Sensor Networks, Vol. 9, No. 1, Article 1, 2012.

[15] P. Lorkowski, T. Brinkhoff, Towards Real-Time Processing of Massive Spatio-Temporally Distributed Sensor Data: A Sequential Strategy Based on Kriging. In: F. Bacao et al. (Eds), AGILE 2015: Geographic Information Sciences as an Enabler of Smarter Cities and Communities, Switzerland: Springer, 2015.

[16] Michael A. Osborne, Stephen J. Roberts, Alex Rogers, Icholas R. Jennings, Real-Time Information Processing of Environmental Sensor Network Data Using Bayesian Gaussian Processes. ACM Transactions on Sensor Networks, Vol. 9, No. 1, Article 1, 2012.

[17] H. Wackernagel, Multivariate Geostatistics: An Introduction with Applications. Springer-Verlag, Berlin Heidelberg, 2003.

[18] A. C. Walkowski, Modellbasierte Optimierung mobiler Geosensornetzwerke für raumzeitvariante Phänomene. Dissertation. Heidelberg: AKA Verlag, 2010.

[19] R. Webster, M. A. Oliver. Geostatistics for Environmental Scientists (Statistics in Practice). West Sussex, England: Wiley, 2007.

[20] H. Wei, Y. Du, F. Liang, C. Zhou, Z. Liu, J. Yi, K. Xu, D. Wu, A k-d tree-based algorithm to parallelize Kriging interpolation of big spatial data. GIScience & Remote Sensing, Vol. 52, No 1, 40-57, Taylor & Francis, 2015.

[21] J.C. Whittier, S. Nittel, M. A. Plummer, Q. Liang, Towards Window Stream Queries Over Continuous Phenomena. 4th ACM SIGSPATIAL International Workshop on GeoStreaming (IWGS) 2013, Orlando, Florida, USA, 2013.

[22] J. Wu, K. Aberer, Y. Zhou, K. L. Tan, Towards Integrated and Efficient Scientific Sensor Data Processing: A Database Approach. 2009. Online Document: http://www.edbt.org/Proceedings/2009-StPetersburg/edbt/papers/p0922-Wu.pdf