Efficient fusion of aggregated historical data

Zongge Liu

1 Abstract

Background. In this paper, we address the challenge of recovering a time sequence of counts from aggregated historical data. For example, given a mixture of the monthly and weekly sums, how can we find the daily counts of people infected with flu? In general, what is the best way to recover historical counts from aggregated, possibly overlapping historical reports, in the presence of missing values? Equally importantly, how much should we trust this reconstruction? Current methods fail to handle complex cases such as missing value, conflicting and overlapping report, while our method not only deal with these cases successfully, but also recover the time sequence with higher accuracy by incorporating domain knowledge.

Aim. In this project, we are particularly interested in this question: how can you recover historical events form aggregated and overlapping historical reports? That is, suppose that we are interested in an unknown time sequence $\vec{x} = \{x_1, x_2, \ldots, x_n\}$ (daily observation of certain event), given several aggregated reports(monthly sums, yearly sums), how can we reconstruct the original sequence from them?

Data. Our dataset is the Tycho dataset, which is a project at the University of Pittsburgh to advance the availability and use of public health data for science and policy making. Currently, the Project Tycho database includes data from all weekly notifiable disease reports for the United States. It dates back to 1888 and covers all the states in US. The types of diseases include measles, smallpox etc.

Method. We provide H-FUSE, a novel method that solves above problems by allowing injection of domain knowledge in a principled way, and turning the task into a well-defined optimization problem utilizing regularization strategies based on knowledge of the historical events, such as smoothness and periodicity. H-FUSE has the following desirable properties: (a) *Effectiveness*, recovering historical data from aggregated reports with high accuracy; (b) *Self-awareness*, providing an assessment of when the recovery is not reliable; (c) *Scalability*, computationally linear on the size of the input data.

Results. Experiments on the real data (epidemiology counts from the Tycho project) demonstrates that H-FUSE reconstructs the original data $30 - 80\%$ better than the least squares method.

Conclusions. We develop a way to recover a time sequence from its partial sums, by formulating it as an optimization problem with various constraints which allows the injection of domain knowledge. Our work extends the previous pseudo-inverse method, and will provide a new way to reconstruct time series from historical data with faster performance and higher accuracy.

Original work published in 2017 SIAM International Conference on Data Mining conference proceeding, published by the Society for Industrial and Applied Mathematics (SIAM).
SDM'17, April 27-29, 2017, Houston, TX, USA

2 Introduction

In this project, we address the challenge of recovering a time sequence of counts from aggregated historical data, which is a part of information fusion problem. The goal of information fusion is to reconstruct objects from multiple resources and observations. Especially, it requires resolving redundancy and inconsistency between observations. This concept has been applied in various domains and under different assumptions. It includes multi-sensor data fusion [10], information fusion for data integration [1], and more recently, human-centered information fusion methods [11].

In modern interdisciplinary research, a comprehensive understanding for the whole picture of the subject requires large amounts of historical data from different data sources acquired by various disciplines. Therefore, the advancement in dealing with information fusion problem is very important for interdisciplinary research. One example would be epidemiological data analysis, which often relies upon knowledge of population dynamics, climate change, migration of biological species, drug development, etc. Another example is the task of exploring long-term and short-term social changes, which requires consolidation of a comprehensive set of data on social-scientific, health, and environmental dynamics, etc. Examples of related applications include monitoring resilient outdoor events and multi-robot search and rescue [23, 24]. Both tasks require large-scale information consolidation from heterogeneous data sources including infrastructure-based mobile systems, ad-hoc wireless networks and distributed Internet repositories.

Nowadays, there are numerous historical data sets available worldwide. For example, the ongoing projects include Great Britain Historical GIS at Portsmouth, the Institute for Quantitative Social Science and the Center for Geographic Analysis at Harvard, the CLIO World Tables at Boston University, the International Institute of Social History in Amsterdam, and World-Historical Dataverse at the University of Pittsburgh etc. Relevant previous projects of data collection and analysis include the Electronic Cultural Atlas Initiative (ECAI), the Integrated Public Use Microdata Series (IPUMS, at Minnesota), the Alexandria Digital Library (ADL) etc. Most notably, in health sciences, the Vaccine Modeling Initiative at the University of Pittsburgh aims to gather and analyze the information from thousands of reports on United States epidemiological data for more than 100 years.

While the aforementioned initiatives indicate a considerable effort to utilize diverse historical data sources, researchers are nowhere near to having a global consolidated historical data repository against which to perform comprehensive socio-scientific analysis and to test emerging large-scale theories. The existing data sources are principally oriented toward regional comparative efforts rather than global applications. They vary widely both in content and format.

3 Problem Statement

Now we summarize the previous discussion into a mathematical problem. In all cases, we have an unknown time sequence of interest $\vec{x} = \{x_1, x_2, \ldots, x_T\}$ (say, of count of measles incidents in New York, per week), and our goal is to reconstruct it, from aggregated information (the sum or weighted sum of the counts).

Problem Definition 1 (Information Fusion)**.** *Informally, the problem is defined as follows:*

- Given: *several (aggregated) reports for the target sequence $\vec{x}$, (for example, some of the monthly sums from source 'A', and some of the yearly sums from source 'B')*
- Reconstruct *the target sequence $\vec{x}$ with modest computational cost (sub-quadratic) and descent accuracy (outperforms the state-of-art methods), and*
- Self-awareness: *how accuracy we can expect the reconstruction to be under different assumptions*

The challenges are the following:

- *Conflicts/Overlaps*: the reports may overlap, or even worse, conflict to each other, e.g., the sum of the monthly reports for some year from source 'A' might **not** be the same as the count for that year, from source 'B'. For example, we may have hundreds of reports from different authorities about cases of measles in Los Angeles in 1900.
- *Missing values*: Due to historical reasons, maybe our reports covering some particular time period, say, 1940-1944, is missing due to the World War II
- *Trust in results*: How confident should people be for the reconstructed values? We will show later that in most cases our proposed method is very good, but in some cases, *no* method can do good reconstruction.

4 Background and Related Work

In this section, we briefly introduce our problem background, and some related works.

4.1 Mathematical Background and Formulation

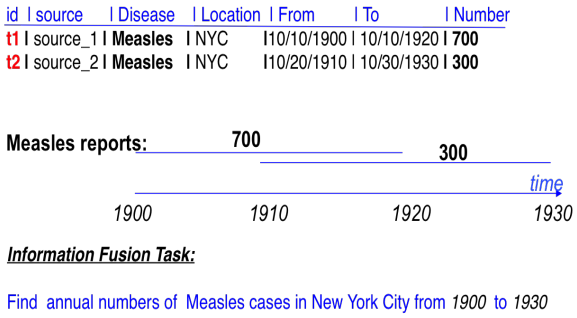


Figure 1: **Illustration on the nature of the data**. Time-overlapping historical reports

A major challenge in historical information fusion is estimating number of historical events from multiple aggregated reports while handling redundant, and possibly, inconsistent information. Figure 1 shows an example of a database with two historical reports on total cases of measles in NYC overlapping in time. Either of the reports are covering time intervals of 20 years. The task of information fusion would be estimating the population dynamics within smaller time units (e.g., what was the most likely annual numbers of measles cases in NYC from 1900 to 1930?). Granularity of the reports may differ. For example, we may need to estimate weekly numbers from monthly reports, or daily values from weekly aggregates. The process of information fusion requires efficient dis-aggregation of the reported data. In general, this problem can be stated in wider context of fusion and making sense of data obtained from a variety of sources, with gaps and overlaps in time and space, and uncertainty in trust of sources.

In our approach, we represent the overlapping historical report as a system of linear equations, – a *characteristic linear system*, as shown in Figure 2. Each report generates a binary row vector for coverage in an observation matrix with "ones" corresponding to the time units covered by the report. The characteristic

system is commonly under-determined and we need to find a reasonably accurate approximate solution that would correspond to the dis-aggregated information.

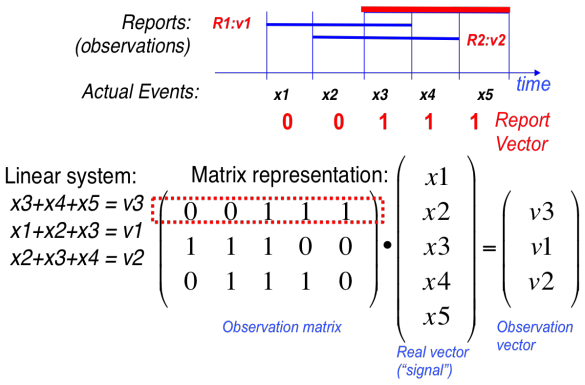


Figure 2: **Illustration on our problem setting**. Characteristic Linear System of time-overlapping historical reports

We call our method 'H-FUSE'

4.2 Related Work

Our work belongs to the general large-scale information fusion problem to deal with many different data sources. A prominent example of a large-scale information fusion project is Tycho [17, 20, 22]. Currently, Tycho collects and consolidates information from approximately 50,000 reports on United States epidemiological data spanning more than 100 years. We used Tycho data for experimental evaluation of our method. Historical information fusion task often involves disaggregation of historical reports, and more mathematically, solving under-determined linear system. Disaggregation methods in time domain have been studied in modern economy research as a subproblem in time series analysis (see [3, 19, 4] for review). The whole process is to reconstruct a high resolution time series from low resolution time series satisfying temporal aggregation constraints, which make the resulting high resolution time series consistent with the low resolution time series in the sense that the sum, average, the median or some specific value of high resolution time series match the low resolution time series.

Related series observed at the required high frequency called indicators can be helpful to dis-aggregate the original observations when they are available to use. However, care must be taken when selecting indicators since two strongly correlated low frequency time series may not be correlated at a higher frequency [7]. Thus, more cares needs to be taken when choosing good indicator series. Temporal dis-aggregation methods have been used for the cases of non-overlapping aggregated reports and cannot be directly applied to the task of historical information fusion.

Conflict resolution is another issue to concern about in information fusion. The current truth recovery algorithms are built on the following principle: reliable sources should provide more reliable information, and reliable information should come more often from reliable truth [14]. Recent work include probabilistic modeling using randomized Gaussian mixture model [21], confidence-aware approach incorporating variance into weight construction [13] etc. However, little work has been done with the historical reports and time series.

Table 1 contrasts our H-FUSE method against the related state-of-the-art competitors. We present our method in the next section.

Property	*LSQ*	*VLDB97[5]*	*TDisaggregation*	*H-FUSE*
Scalability	✓	✓		✓
Self-awareness		✓		✓
Overlapping reports	✓	?		✓
Missing values		?		✓
Conflicting reports				✓
Domain knowledge		?	✓	✓

Table 1: **H-FUSE matches all specs**, while competitors miss one or more of the features.

The challenge of data fusion has been studied in various domains such as image processing, signal processing, geology, etc, that deals with uncertainty in the task. Image and signal processing community has been dealing with related ill-posed problems such as edge detection, surface reconstruction [25], vehicle detection [18], or super-resolution image reconstruction [15, 16]. Similarly, in geology researchers need to handle tomographic pumping tests [2] or mineral mining [9]. Common methods to derive an approximate solution for an under-determined linear system include the least squares method (LSQ) and Tikhonov regularization method[8, 6], by introducing additional constraints such as smoothness in space and/or temporal domain to in-cooperate the domain knowledge for the data. Although Tikhonov regularization has been widely applied to solve ill-posed problem in various research fields, to our knowledge the application of Tikhonov regularization has not been addressed in the historical information fusion context.
In 'VLDB97', Faloutsos et. al. [5] uses smoothness, for information fusion. The summaries were expected to be consecutive and non-overlapping; we believe that it could handle overlaps and missing values with proper extension (indicated as '?' in Table 1). However, the method is completely incapable of handling conflicts.

5 Method

In this section, we explain our H-FUSE in more details. Table 2 gives the list of symbols we use.
The common requirement in all reconstruction methods, is that the reconstructed sequence $\vec{x}$ should satisfy the reports/facts, that is

$$\mathcal{F}(\vec{x}) = \sum_{n=1}^{N} (v_n - \sum_{t=1}^{T} \mathbf{O}_{nt} x_t)^2 \tag{1}$$

and, in matrix form:

$$\mathcal{F}(\vec{x}) = ||\vec{v} - \mathbf{O}\vec{x}||_2^2 \tag{2}$$

Ideally, the deviation from the facts should be zero, unless the facts/reports are conflicting. The top competitor, 'LSQ', stopped here, and tried to minimize $\mathcal{F}()$; since the problem is (usually) under-determined,

'LSQ' proposed to find the minimum-norm solution ($\min\ ||\vec{x}||_2^2$) that satisfies $\mathcal{F}$. This is a well-understood problem, and 'LSQ' can find that unique solution using the so-called *Moore-Penrose pseudo-inverse*.
But there is no reason why the solution would have minimum norm, which leads to our proposed solution.

Symbols	Definitions
$\vec{x}$	$= (x_1, \ldots, x_T)$: target time series (*unknown.*)
T	total number of timeticks in $\vec{x}$
P	(smallest) period of $\vec{x}$
N	total number of reports
D	report duration
$\vec{v}$	$= (v_1, \ldots, v_N)$: values of reports (*observed* - aggregated form of *unknown* $\vec{x}$)
$\mathbf{O}$	$N \times T$ observation matrix ($\vec{v} = \mathbf{O}\vec{x}$)
$\mathcal{F}(\vec{x})$	deviation from facts/reports
$\mathcal{C}(\vec{x})$	domain-imposed soft constraint
$\mathcal{L}(\vec{x})$	total penalty ('loss')
$\mathbf{H}_s$	$(T-1) \times T$ smoothness matrix
$\mathbf{H}_p$	$(T-P) \times T$ periodicity matrix

Table 2: Symbols and Definitions

5.1 Intuition

The main idea behind our H-FUSE is to infuse domain knowledge. For example, in most cases where the solution sequence $\vec{x}$ should be smooth, we propose to penalize large differences between adjacent timeticks; if we know that the periodic, we propose to also impose periodicity constraints.
More formally, our approach is to find the values (x_t, $t = 1, \ldots, T$) that (a) can be aggregated to generate observed report ($\vec{v}$) and (b) minimize a domain-dependent penalty functions. Thus, we propose to formulate the optimization problem as follows:

$$\min_{\vec{x}} \mathcal{L}(\vec{x}) = \min_{\vec{x}}\ (\mathcal{F}(\vec{x}) + \mathcal{C}(\vec{x})) \tag{3}$$

where $\mathcal{L}(\vec{x})$ stands for the total penalty ('loss', hence the symbol $\mathcal{L}$), and consists of two components: The first is $\mathcal{F}(\vec{x})$, the deviation from the reports ('facts'), that we defined before (Eq. 2). The second component, $\mathcal{C}(\vec{x})$, infuses domain knowledge, in the form of soft constraints, like smoothness and periodicity, that we explain below. It could also infuse other types of domain knowledge, like sparsity, adherence to an epidemiology model like SIS (susceptible-infected-susceptible, like the flu), but we will not elaborate here. Let us focus on the two constraints that we propose, since they proved to be the most successful in our experiments.

- Smoothness constraint $\mathcal{C}_s$ This constraint penalizes big jumps between successive timeticks. Formally:

$$\mathcal{C}_s(\vec{x}) = \sum_{t=1}^{T}(x_t - x_{t+1})^2 = ||\mathbf{H}_s\vec{x}||_2^2 \tag{4}$$

where $\mathbf{H}_s$ is a $\mathbb{R}^{(T-1)\times T}$ matrix whose t^{th} row has 1 and -1 in the t^{th} and $(t+1)^{th}$ column, respectively.

- Periodicity constraint $\mathcal{C}_p$: If there is a period (say P=52 weeks = 1year) in our data, we can penalize deviations from that, as follows:

$$\mathcal{C}_p(\vec{x}) = \sum_{t=1}^{T}(x_t - x_{t+P})^2 = ||\mathbf{H}_p\vec{x}||_2^2 \tag{5}$$

where $\mathbf{H}_p$ is a $\mathbb{R}^{(T-P)\times T}$ matrix whose t^{th} row has 1 and -1 in the t^{th} and $(t+P)^{th}$ column respectively. The intuition here is to make the event at timetick t to be close to the one at $t+P$.

5.2 Problem Design

As briefly mentioned in previous section, the main optimization function is the reconstruction function shown in Equation 3. Given N observed reports $\vec{v}$, we are trying to recover the target time series vector $\vec{x}$ that is aggregated in various time periods to generate the observed reports.
The target time series $\vec{x}$ is designed to consists of time sequence of events that are in equi-space intervals. The aggregation of the target time series $\vec{x}$ to $\vec{v}$ is done by multiplying the observation matrix $\mathbf{O}$ where each row is responsible for generating one report by selective addition of the elements in $\vec{x}$. For example, if the timetick of x corresponds to year $(1970, 1980, 1990, 2000, \cdots)$, and we have a report v_n that is sum of events over period $1970 - 1990$, then corresponding row of $\mathbf{O}$ matrix will be $(1, 1, 1, 0, \cdots)$.
Subtle issue: relative weights: A careful reader may wonder whether we should give different weight to deviations from the facts $\mathcal{F}(\cdot)$, as opposed to the (soft) constraints/conjectures $\mathcal{C}(\cdot)$.
The short answer is '*no*'. The long answer is that we tried a weighting parameter λ, and we tried to minimize the loss function $\mathcal{L}(\cdot) = \mathcal{F}(\cdot) + \lambda\mathcal{C}(\cdot)$. However, we recommend to set λ=1, for the following reasons: (a) it gives optimal, or near-optimal results, for all real cases we tried (as compared to the ground truth $\vec{x}$); (b) the results are insensitive to the exact value of λ, when $\lambda \leq 1$; (c) it is hard for a practitioner to set the value of λ, given that the target sequence is unknown.

5.3 Methods and Procedures

In short, smoothness is the constraint that we propose to use as default, if the domain expert has nothing else to tell us. The reason is that it performs well, as we see in the experiments, as well as for theoretical reasons. If the domain expert believes that there is a periodicity with period P, then we can do even better. The exact problem formulations are as follows.

5.3.1 H-FUSE-S (Smoothness Method)

The proposed loss function $\mathcal{L}_s(\cdot)$ is given by Equation 3

$$\mathcal{L}_s(\vec{x}) = \mathcal{F}(\vec{x}) + \mathcal{C}_s(\vec{x})$$

which gives, in matrix form:

$$\min_{\vec{x}} \mathcal{L}_s(\vec{x}) = \min_{\vec{x}} \ (||\vec{v} - \mathbf{O}\vec{x}||_2^2 + ||\mathbf{H}_s\vec{x}||_2^2) \tag{6}$$

5.3.2 H-FUSE-P (Periodicity Method)

Time sequence data are often periodic. Historical events such as epidemics, weather measurements demonstrate repeating cycles of patterns. For example, flu outbreak records may have a seasonal pattern where there is a peak in the winter and recovery in the summer season. Weather measurements may demonstrate cyclic patterns in days, and seasons.
For such cases, we propose to impose both smoothness, as well as periodicity constraints, with the period P that a domain expert will provide.
Then, we propose the loss function $\mathcal{L}_p(\cdot)$ to be

$$\mathcal{L}_p(\vec{x}) = \mathcal{F}(\vec{x}) + 1/2\mathcal{C}_s(\vec{x}) + 1/2\mathcal{C}_p(\vec{x})$$

which leads to the optimization problem below:

$$\min_{\vec{x}} ||\vec{v} - \mathbf{O}\vec{x}||_2^2 + 1/2||\mathbf{H}_s\vec{x}||_2^2 + 1/2||\mathbf{H}_p\vec{x}||_2^2 \tag{7}$$

We chose equal weights of $1/2$ for each of the constraints, so that they will not overwhelm the facts-penalties $\mathcal{F}(\cdot)$. In any case, the reconstruction quality is rather insensitive to the exact choice of weights, with similar arguments as we discussed earlier about the λ weight (see subsection 5.1, page 7).

6 Analysis and Result

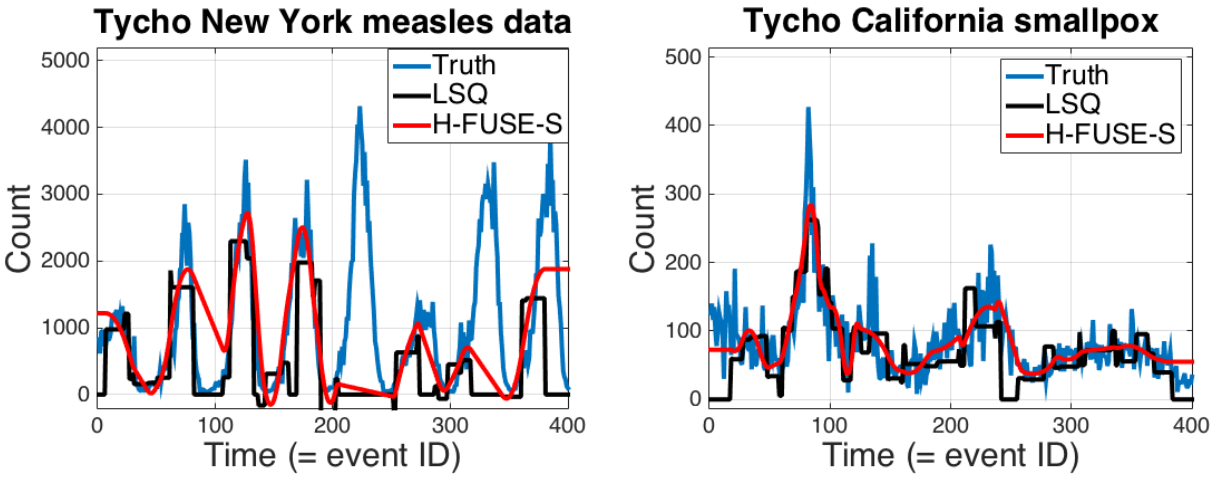
In this section we report experimental results of our H-FUSE on the real data.

6.1 Experimental setup and Analysis

To prove the effectiveness of H-FUSE on the real data, we apply H-FUSE on Tycho [17] New York measles data which is our default main dataset. The full dataset contains 3952 weekly records with some missing values. We carefully select the time period without missing values, Week 51 to Week 450, which gives us 400 records in total as our selected dataset. To test the generality of H-FUSE, we also select 400 weekly records from California smallpox data, ranging from Week 501 to Week 900.
We refer to the observations or records as *reports*, and the number of observations as *report numbers*, and the timeticks that each report covers as *report duration*. We vary the report numbers and the report duration to conduct sensitivity test of H-FUSE. The reports are generated randomly for specific report number and report duration combination, i.e., the mixing matrix $\mathbf{O}$ is constructed to reflect the report number and duration that we set. We then apply our method into these simulated data.

6.2 Effectiveness of H-FUSE-S

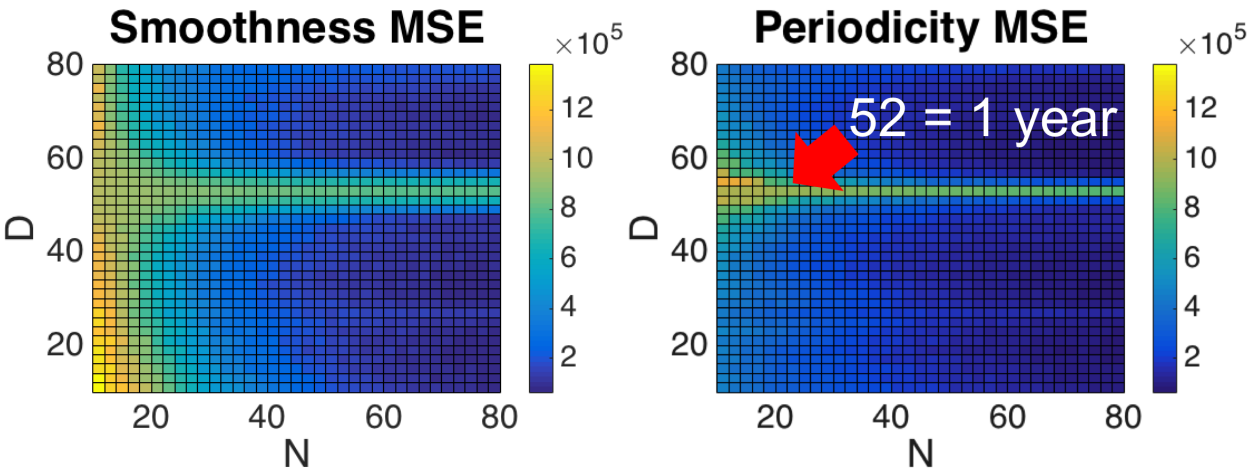
In this section, we compare the reconstruction performance of H-FUSE-S and the conventional approach, LSQ method. In Figure 3, the reconstruction comparisons of LSQ method and H-FUSE-S are shown for (a) Tycho New York measles data, and (b) Tycho California smallpox data. We see that generally, H-FUSE with smoothness constraint gives better reconstruction than the LSQ method.
We further conducted experiment under various configurations of report number and report duration ranging from 10 to 80. For each configuration, we repeat the experiment 100 times, and average over the reconstruction MSE. The error dynamics of H-FUSE with smoothness constraint on various settings of report number and report duration is shown in Figure 4 (a). We vary the report number and the report duration in the x-axis



(a) Tycho New York measles data - 30% improvement over LSQ (b) Tycho California smallpox data - 58% improvement over LSQ

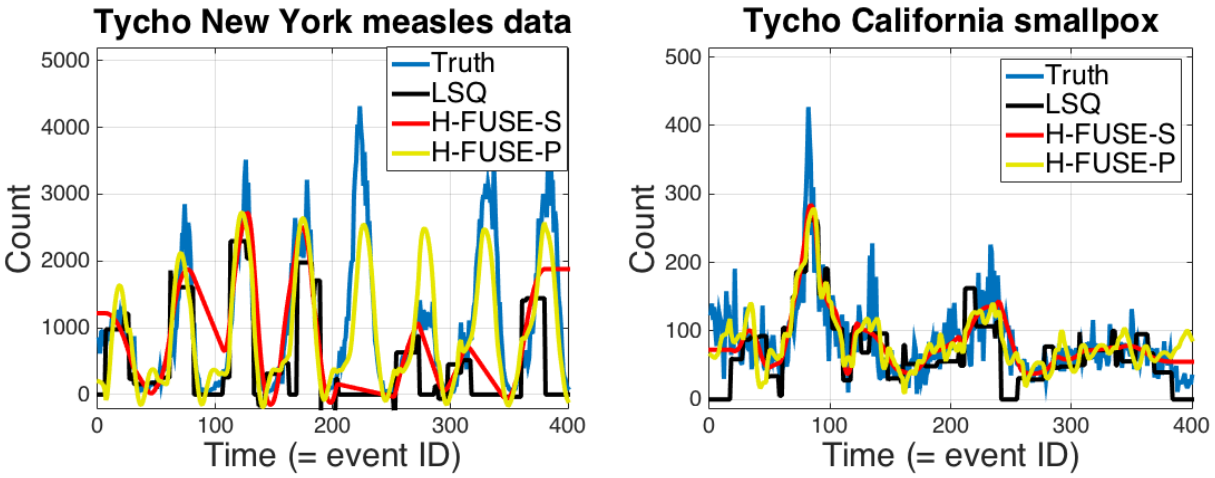
Figure 3: **H-FUSE-S reconstructs well** H-FUSE-S wins over LSQ method in reconstruction all along. (a) Tycho New York measles data ($N = 20, D = 20$) - H-FUSE-S reconstructs 30% better than LSQ. (b) Tycho California smallpox data ($N = 30, D = 30$) - H-FUSE-S reconstructs 58% better than LSQ.

and y-axis, respectively. Here brighter color indicates higher MSE in reconstruction. From the figure, we observe a clear trend of decrease in reconstruction MSE as the report number increases, and MSE reach its minimum for a combination of large report number and long report duration.



(a) H-FUSE-S (b) H-FUSE-P

Figure 4: **H-FUSE with various configurations**. Both versions of H-FUSE reconstructs well (*blue*) unless the report duration matches exactly with the period $D = P(= 52)$.



(a) Tycho New York measles data - 30%, 81% improvement by H-FUSE-S, H-FUSE-P over LSQ

(b) Tycho California smallpox data - 58%, 41% improvement by H-FUSE-S, H-FUSE-P over LSQ

Figure 5: **H-FUSE *is effective***: Tycho [17] (a) New York measles and (b) California smallpox counts per week (in *blue*). H-FUSE (in *red* and *yellow*) captures the cycles, outperforming top competitor 'LSQ' (in *black*).

6.3 Effectiveness of H-FUSE-P

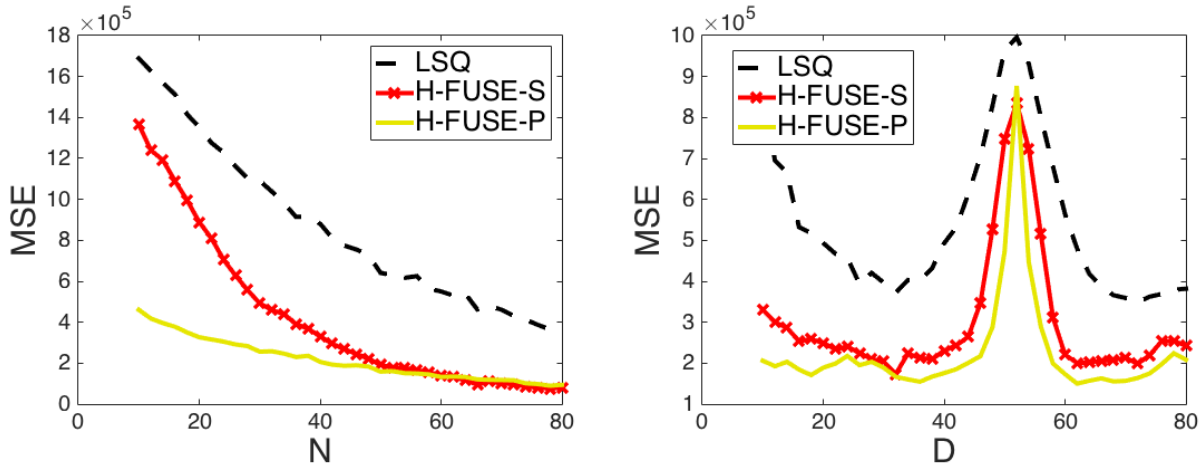
In the case of Tycho New York measles data, it is known that it has a periodicity of one year. Therefore, we apply $\mathcal{C}_p$ periodicity constraint in addition to $\mathcal{C}_s$ smoothness constraint on the measles data, i.e. H-FUSE-P as described in Section 5.3.2. In Figure 5, it plots the New York measles data (in blue, 'Truth'), the reconstruction of the top competitor (in black, 'LSQ'), and the two versions of H-FUSE ('Smoothness' in red, and 'Periodicity' in yellow). Our reconstructions are visibly better than 'LSQ', with up to 80% better reconstruction. The reconstruction comparisons of LSQ method, H-FUSE-S and H-FUSE-P give us a clear picture on how periodicity constraint improves the performance in addition to that of the smoothness constraint. We observe that in almost all cases the additional periodicity constraint results in smaller MSE than the smoothness constraint alone.

The error dynamics of H-FUSE-P on various settings of report number and report duration is shown in Figure 4 (b). We observe a similar trend as in the simple smoothness constraint case: the MSE decreases as the report number increases. However, we observe that the additional periodicity constraint significantly improves the accuracy in terms of MSE.

In fact, among all of the report number and report duration configurations, there were only a very few case when H-FUSE-S outperformed the H-FUSE-P. The result is illustrated in more detail in Figure 4 (b). These cases arise because when report number is high, neither assumption is required to be strong any more. Therefore they have similar level of performance.

In Figure 6 (a), we study the change in MSE with varying report numbers. The overall observation is that the MSE decreases with the increase of the number of reports.

We also analyzed the change in MSE with varying report duration in Figure 6 (b). We observe that MSE displays a periodic pattern, reaching its peaks when $D = 52$. The reason is that when report duration matches the periodicity of the data, all of the values in any of the report will have similar values, leading to deficit of information.



(a) error vs. report number

(b) error vs. report duration

Figure 6: **H-FUSE wins consistently** (a) Error decreases consistently with report number N. (here $D = 40$) (b) Error is almost constant with respect to report duration D unless $D = P = 52$ (here $N = 40$).

7 Theory and Discussion

In this section we provide theoretical analysis on H-FUSE. From the theoretical analysis, we induce a set of observations that provide an assessment of the cases when the reconstruction is not reliable, which we refer to as "self-awareness" of H-FUSE. Also, we demonstrate the scalability property of H-FUSE in terms of both theory and empirical aspects.

7.1 Background

Consider

$$\min_{\vec{x}} ||\vec{v} - \mathbf{O}\vec{x}||_2^2 + ||\mathbf{H}\vec{x}||_2^2 \tag{8}$$

which is equivalent to

$$\min_{\vec{x}} \left|\left| \begin{bmatrix} \vec{v} \\ - \\ \vec{0} \end{bmatrix} - \begin{bmatrix} \mathbf{O} \\ - \\ \mathbf{H} \end{bmatrix} \vec{x} \right|\right|_2^2. \tag{9}$$

Here $\mathbf{H}$ is of type

$$\mathbf{H}_s = \begin{bmatrix} -1 & 1 & 0 & 0 & 0 & 0 \\ 0 & -1 & 1 & 0 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 & 0 \\ 0 & 0 & 0 & -1 & 1 & 0 \\ 0 & 0 & 0 & 0 & -1 & 1 \end{bmatrix}, \tag{10}$$

Signals in the right null space of the matrix is smooth. In the case of $\mathbf{H}_s$, the signals are constant. More generally, notice that if $|\vec{x}(n) - \vec{x}(n-1)| \leq \epsilon$, then $||\mathbf{H}_s\vec{x}||_2^2 \leq \epsilon^2(T-1)$, where $T = \text{length}(\vec{x})$.
In short, the inclusion of the regularizer $||\mathbf{H}\vec{x}||_2^2$, nudges $\vec{x}$ into a smoother solution vector.

7.2 Self-awareness

What can we say about the error of reconstruction? We give the theoretical analysis here, and later on, show how they translate to practical recommendations.
The first lemma states that if our target sequence $\vec{x}$ satisfies a smoothness condition, and if we have enough 'suitable' equations, then we can have error-free reconstruction.
Formally, let $\mathbf{O}$ denote the $N \times T$ observation matrix, $\vec{x}$ be the target time series, and $\mathbf{H}$ is one of the smoothness regularization matrices in Eq. 10.

Lemma 1. *With the above notations, if (a)* $\begin{bmatrix} \mathbf{O} \\ - \\ \mathbf{H} \end{bmatrix}$ *is tall or square and full column rank, and (b)* $\mathbf{H}\vec{x} = 0$*, then we can have error-free reconstruction.*

Proof. Special case of the upcoming Lemma 2. □

The first condition (full column rank) is almost always true; the second condition is rather strict. The next Lemma relaxes it, effectively stating that if our target sequence is close to smooth ($\mathbf{H}\vec{x} \approx 0$), then the reconstruction error is small. Formally, with the same notations as above, we have:

Lemma 2. *If* $\begin{bmatrix} \mathbf{O} \\ - \\ \mathbf{H} \end{bmatrix}$ *is tall or square and full column rank, the squared error for our smoothness reconstruction is given by*

$$SE = ||(\mathbf{O}^T\mathbf{O} + \mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{H}\vec{x}||_2^2 \tag{11}$$

Proof. Let $\hat{\vec{x}}$ be the solution we obtain from solving Equation 8. This is an over-determined system (more rows/equations than unknowns/columns), and the solution is given by

$$\hat{\vec{x}} = (\mathbf{O}^T\mathbf{O} + \mathbf{H}^T\mathbf{H})^{-1}\mathbf{O}^T\vec{v}$$

Then for $\vec{x}$, we have trivially that

$$\vec{x} = \mathbf{I}\vec{x}$$

with $\mathbf{I}$ being the $T \times T$ identity matrix. Then:

$$\begin{aligned}\vec{x} &= \left(\begin{bmatrix} \mathbf{O}^T | \mathbf{H}^T \end{bmatrix} \begin{bmatrix} \mathbf{O} \\ - \\ \mathbf{H} \end{bmatrix} \right)^{-1} \begin{bmatrix} \mathbf{O}^T | \mathbf{H}^T \end{bmatrix} \begin{bmatrix} \mathbf{O}\vec{x} \\ - \\ \mathbf{H}\vec{x} \end{bmatrix} \\ &= (\mathbf{O}^T\mathbf{O} + \mathbf{H}^T\mathbf{H})^{-1}(\mathbf{O}^T\vec{v} + \mathbf{H}^T\mathbf{H}\vec{x}) = \hat{\vec{x}} + \left(\mathbf{O}^T\mathbf{O} + \mathbf{H}^T\mathbf{H}\right)^{-1}\mathbf{H}^T\mathbf{H}\vec{x}\end{aligned}$$

Thus, the error vector $\vec{e} = \vec{x} - \hat{\vec{x}}$ is

$$\vec{e} = \left(\mathbf{O}^T\mathbf{O} + \mathbf{H}^T\mathbf{H}\right)^{-1}\mathbf{H}^T\mathbf{H}\vec{x}$$

and its squared norm is given by Eq. 11 □

Clearly, when $\mathbf{H}\vec{x} = 0$, the loss becomes zero, which is exactly what Lemma 1 says. Otherwise, the error depends on how much $\vec{x}$ deviates from smoothness ($\mathbf{H}\vec{x}$), as well as the specifics of the $\mathbf{O}$ and $\mathbf{H}$ matrices. The above lemmas hold for arbitrary starting points of the reports, and for arbitrary lengths. We can provide additional bounds and guide-lines, for the (very realistic) setting that the reports all have the same length D. We distinguish 3 settings, in this case:

- *semi-random report*: The starting points of the reports, are random. Thus, they may overlap, and/or coincide, and/or leave uncovered parts of the target signal $\vec{x}$.
- *tile report*: The reports have deterministic starting points (1, D+1, $2 * D$+1, ...) and thus they cover the whole time interval, without overlaps.
- *shingle report*: General case of *tile report*: successive reports overlap by o time-ticks (for *tile report*, o=0).

Then we can give additional guarantees.

Lemma 3. *Given an infinite target time series $\vec{x}$, with smallest period P; in a* tile report *setting of duration D, if*

$$D < P/2$$

then there exists a method for reconstructing the signal with no error.

Proof. See [5]. The proof is closely related to the Nyquist sampling frequency. □

Lemma 4. *Given an infinite target time series $\vec{x}$, with smallest period P in a* shingle report *setting of duration D, if*

$$D < P/2$$

then there exists a method for reconstructing the signal with no error.

Proof. (Sketch:) Choose the subset of reports that form a *tile report* setting - by Lemma 3 we can have error-free reconstruction. □

Informally, the above Lemmas explains our empirical observations which show that if the target time series $\vec{x}$ is finite and periodic with smallest period P (52 weeks, in our measles data), and if we have *tile report* (or *shingle report*) reports with $D < P/2$, H-FUSE (with its smoothness constraint) will result in small error. When the $D \geq P/2$, there are no recommendations any more - H-FUSE may, or may not, result in large errors.

7.3 Scalability

Our H-FUSE eventually needs to solve a sparse linear system, and, intuitively, this should be fast. This is indeed the case, as we show next. Let D_{max} be the duration of the longest report, and assume that $D_{max} \geq b$, where b is the bandwidth of the $\mathbf{H}$ matrix - b=2 for $\mathbf{H}_s$ as in Section 5.3.1. With the usual notation ($\vec{x}$ is the target sequence, of length T), we have the Lemma:

Lemma 5. *For any report setting, let D_{max} be the longest report duration. Then the total computation time for our* H-FUSE-S *is*

$$O(Tlog(T) + 4D_{max}^2)$$

Proof. The most time consuming part is the matrix inversion. The $\mathbf{O}$ is a banded Toeplitz matrix with bandwidth D_{max}, thus $\mathbf{O}^T\mathbf{O}$ is likewise a banded Toeplitz matrix of bandwidth $2D_{max} - 1$; and the same holds for $\mathbf{H}$ and $\mathbf{H}^T\mathbf{H}$, with bandwidth $b \leq D_{max}$. Then, the result follows from [12]. □

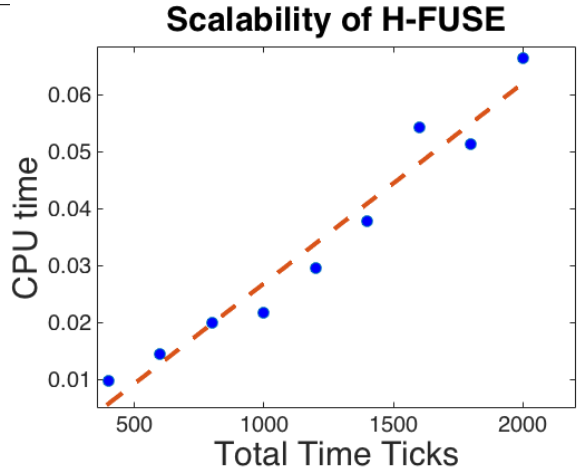


Figure 7: **H-FUSE is scalable** H-FUSE scales near-linearly on the length T. CPU time (seconds) vs T, in lin-lin scales; dashed orange line is plotted for reference.

Empirically, our H-FUSE seems to have even better scalability, close to linear: Figure 7 shows the wall-clock time versus the sequence length T. We used "regular" reports, with duration $D = 200$, on the full New York measles data with $T = 3,952$ time ticks and applied H-FUSE with the smoothness constraint. H-FUSE scales linearly, which is even better than what Lemma 5 predicts.

7.4 Limitation

The limitation of our methods may be summarized as follows,

- Our mothod is still restricted in the linear domain, while in practice information fusion may be stated as a possibly nonlinear problem under certain case.
- Our proposal for $\mathbf{H}$ is too simple, only considering purely smooth or purely periodic ideal cases, while in practice the domain knowledge may appear in more complex form, such as SIR model in epidemic study.

8 Practitioner's Guide

From the above discussion, smoothness has desirable theoretical properties, and, as the experiments show, it is a good choice for solving the information fusion problem. Moreover, if the domain expert knows that there is a periodicity of period, say P, our H-FUSE can incorporate it and achieve even better reconstruction. The question is when should the domain expert trust (or discard) the results of our reconstruction? We summarize our recommendations, next.

Recommendation 1 (Smoothness is effective)**.** *If the target sequence $\vec{x}$ is smooth, and we have enough reports, then* H-FUSE *achieves good reconstruction.*

The error is given by Lemma 2, and it is zero, if $\vec{x}$ is perfectly smooth (Lemma 1) Figure 6 (a) provides evidence.

Recommendation 2 (Nyquist-like setting)**.** *When we have regular reports frequently enough (i.e., with $D < P/2$), then we can expect small error from our* H-FUSE*.*

This is the informal version of Lemma 3, and illustrated in Figure 6. P is the smallest period of our target signal (eg., P=52 weeks, in our measles data).
Finally, we did not provide a proof, but the recommendation is obvious. Given reports of the same length D, we have:

Recommendation 3 (Obliteration)**.** *If the report length D coincides with the period P of the signal, large errors are possible.*

The intuition is that, say, if all our reports span exactly D=52 weeks (=1 year), there is no way anyone can recover the annual (March) spikes of measles. Figure 4 gives us an arithmetic example where you have large MSE for D=52 weeks.

9 Conclusion

We proposed H-FUSE method that efficiently reconstructs historical counts from possibly overlapping aggregated reports. We propose a principled way of recovering a times sequence from its partial sums, by formulating it as an optimization problem with various constraints (Eq. 3). Our formulation allows the injection of domain knowledge (smoothness, periodicity, etc). Our method has the following major properties:

1. **Effectiveness**: The experimental result on the real-world Tycho New York measles data, outperformed the reconstruction by naive approach as shown in Section 6.
2. **Self-awareness**: We provide theoretical results that help evaluate the quality of the reconstruction as discussed in Section 7.2.
3. **Scalability**: The computational cost for H-FUSE scales nicely as $O(Tlog(T) + 4D^2)$ as discussed in Section 7.3.

10 Acknowledgment

Zongge would like to thank the support from the statistics department at Carnegie Mellon University. This material is based upon worksupported by the National Science Foundation under Grants No. IIS-1247489, IIS-1247632. This work is also partially supported by an IBM Faculty Award and a Google Focused Research Award. We would also thank our collaborators outside the committee, V. Zadorozhny and N. Sidiropoulos. V. Zadorozhny was partially supported by NSF BCS-1244672 grant. N. Sidiropoulos was partially supported by NSF IIS-1247632 and IIS-1447788. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation, or other funding parties. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation, or other funding parties. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

Keywords: Information fusion, optimization

DAP Committee members:

Christos Faloutsos $\langle christos + @cs.cmu.edu \rangle$ (School of Computer Science, Carnegie Mellon University);

Robert E. Kass $\langle kass@andrew.cmu.edu \rangle$ (Department of Statistics, Carnegie Mellon University);

Approved by DAP Committee Chair: _____

11 References

References

[1] J. Bleiholder and F. Naumann. Data fusion. *ACM Computing Surveys*, 41(1), 2008.

[2] Geoffrey C Bohling. Information fusion in regularized inversion of tomographic pumping tests. In *Quantitative Information Fusion for Hydrological Sciences*, pages 137–162. Springer, 2008.

[3] I. Brown. An empirical comparison of benchmarking methods for economic stock time series. *US Census Bureau*, 2012.

[4] B. Chen. An empirical comparison of methods for temporal distribution and interpolation at the national accounts. *Bureau of Economic Analysis*, 2007.

[5] Christos Faloutsos, H. V. Jagadish, and Nikolaos Sidiropoulos. Recovering information from summary data. In *VLDB'97, August 25-29, 1997, Athens, Greece*, pages 36–45, 1997.

[6] Silvia Gazzola and James G. Nagy. Generalized arnoldi–tikhonov method for sparse reconstruction. *SIAM Journal on Scientific Computing*, 36(2):B225–B247, 2014.

[7] G.Chamberlin. Temporal disaggregation. *Economic and Labour Market Review*, 2010.

[8] Gene H. Golub, Per Christian Hansen, and Dianne P. O'Leary. Tikhonov regularization and total least squares. *SIAM Journal on Matrix Analysis and Applications*, 21(1):185–194, 1999.

[9] Justin Granek and Eldad Haber. Data mining for real mining: A robust algorithm for prospectivity mapping with uncertainties. In *Proceedings of the 2015 SIAM International Conference on Data Mining*, pages 145–153. SIAM, 2015.

[10] D. Hall. *Mathematical techniques in multi-sensor data fusion*. Artech House, 2004.

[11] D. Hall and J. Jordan. *Human-centered information fusion*. Artech House, 2010.

[12] A. Jain. Fast inversion of banded toeplitz matrices by circular decompositions. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26(2):121–126, Apr 1978.

[13] Qi Li, Yaliang Li, Jing Gao, Lu Su, Bo Zhao, Murat Demirbas, Wei Fan, and Jiawei Han. A confidence-aware approach for truth discovery on long-tail data. *Proc. VLDB Endow.*, 8(4):425–436, December 2014.

[14] Yaliang Li, Jing Gao, Chuishi Meng, Qi Li, Lu Su, Bo Zhao, Wei Fan, and Jiawei Han. A survey on truth discovery. *SIGKDD Explor. Newsl.*, 17(2):1–16, February 2016.

[15] Antigoni Panagiotopoulou and Vassilis Anastassopoulos. Super-resolution image reconstruction techniques: Trade-offs between the data-fidelity and regularization terms. *Information Fusion*, 13(3):185–195, 2012.

[16] Vorapoj Patanavijit and Somchai Jitapunkul. A lorentzian stochastic estimation for a robust iterative multiframe super-resolution reconstruction with lorentzian-tikhonov regularization. *EURASIP Journal on Advances in Signal Processing*, 2007(1):1–21, 2007.

[17] Tycho Project:. https://www.tycho.pitt.edu.

[18] R Alberto Salinas, Christopher Richardson, Mongi A Abidi, and Ralph C Gonzalez. Data fusion: Color edge detection and surface reconstruction through regularization. *IEEE Transactions on Industrial Electronics*, 43(3):355–363, 1996.

[19] C. Sax and P. Steiner. Temporal disaggregation of time series. *The R Journal*, 41(5), 2013.

[20] W. van Panhuis, J. Grefenstette, S. Jung, N. Chok, A. Cross, H. Eng, B. Lee, V. Zadorozhny, S. Brown, D. Cummings, and D. Burke. Contagious diseases in the united states from 1888 to the present. *The New England Journal of Medicine*, 369(22), 2013.

[21] Houping Xiao, Jing Gao, Zhaoran Wang, Shiyu Wang, Lu Su, and Han Liu. A truth discovery approach with theoretical guarantee. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 1925–1934, New York, NY, USA, 2016. ACM.

[22] V. Zadorozhny and Y.-F. Hsu. Conflict-aware fusion of historical data. In *Proc. of the 5th International Conference on Scalable Uncertainty Management (SUM11)*, 2013.

[23] V. Zadorozhny, P.-J. Lee, and M. Lewis. Collaborative information sensemaking for search and rescue missions. In *Proc. of the 12h International Conference on Information Systems for Crisis Response and Management (ISCRAM15)*, 2015.

[24] V. Zadorozhny and M. Lewis. Information fusion for usar operations based on crowdsourcing. In *Proceedings of the 16th International Conference on Information Fusion (FUSION13)*, 2013.

[25] Ying Zhu, Dorin Comaniciu, Martin Pellkofer, and Thorsten Koehler. Reliable detection of overtaking vehicles using robust information fusion. *IEEE Transactions on Intelligent Transportation Systems*, 7(4):401–414, 2006.