H-FUSE: Efficient Fusion of Aggregated Historical Data

Zongge Liu* Hyun Ah Song* Vladimir Zadorozhny† Christos Faloutsos *
Nicholas Sidiropoulos‡

Abstract

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?

We propose 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. 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.

Experiments on the real data (epidemiology counts from the Tycho project [13]) demonstrates that H-FUSE reconstructs the original data $30 - 81\%$ better than the least squares method.

1 Introduction

Information fusion is the process of reconstructing objects from multiple observations. The need for information fusion appears in numerous domains, including multi-sensor data fusion [8], information fusion for data integration [1], and human-centered information fusion [9]. It is particularly important for interdisciplinary research, where a comprehensive picture of the subject requires large amounts of historical data from disparate data sources from a variety of disciplines. For example, epidemiological data analysis often relies upon knowledge of population dynamics, climate change, migration of biological species, drug development, etc. Similarly, information fusion is vital for exploring long-term and short-term social changes, where we need to consolidate data on social-scientific, health, and environmental dynamics. Such historical data sets are available from numerous groups worldwide such as the Institute for Quantitative Social Science and the Center for Geographic Analysis at Harvard, or World-Historical Database at the University of Pittsburgh.

In all cases, we have an unknown, target time sequence $\vec{x} = \{x_1, x_2, \ldots, x_T\}$ (say, of count of measles incidents in the USA, per week), and we want to reconstruct it, from aggregated information.

INFORMAL PROBLEM 1. (INFORMATION FUSION) *Informally, the problem is as follows:*

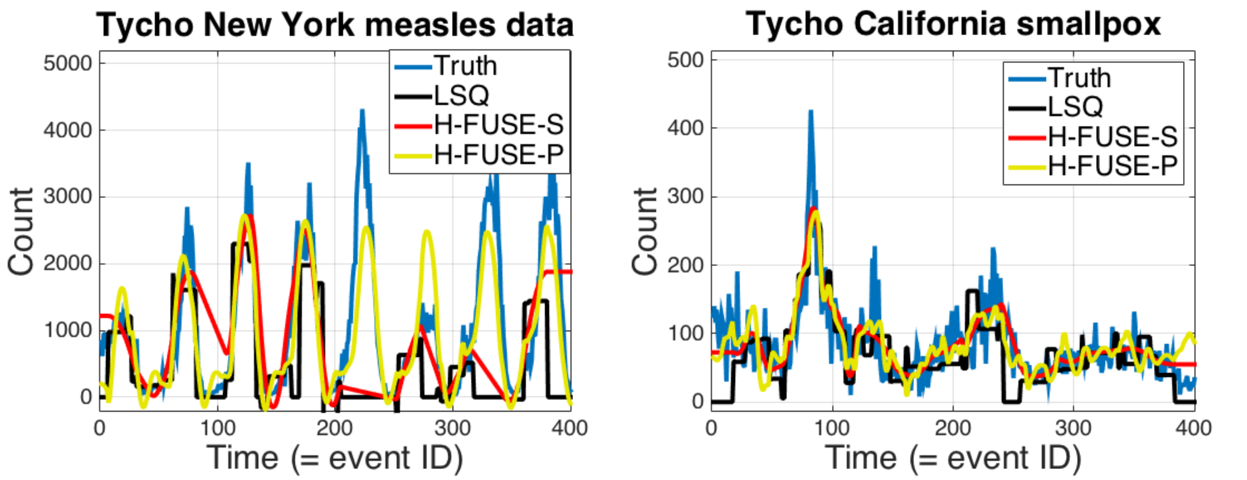
- Given: *several (aggregate) reports for the target sequence $\vec{x}$, (for example, some of the monthly sums from source 'A', and some of the yearly sums from source 'B')*
- Reconstruct *the target sequence $\vec{x}$, as accurately as possible, and*
- Confidence: *indicate whether we should trust the reconstruction.*

The challenges are the following:

- *Conflicts/Overlaps*: the reports may overlap; and even worse, conflict, e.g., the sum of the monthly reports for year, say 1993, might **not** be the same as the count for that year, from source 'B'.
- *Missing values*: Maybe none of our sources ('A' and 'B', above), covers the period of 1940-1944 (because, say, of World War II).
- *Trust in results*: How confident should one be on the reconstructed values? As we show later, in most cases our proposed method is very good, but in some cases, *no* method can do good reconstruction (see Recommendation 3 in Section 6).

We introduce H-FUSE, a systematic and efficient approach to address the problem of fusing aggregated historical data sources. Our H-FUSE consistently disaggregates historical data reports, addressing all the challenges mentioned above. The main idea is to take into account domain knowledge (e.g., smoothness, periodicity) and to infuse it as constraints in an optimization problem, when merging several historical reports.

*Carnegie Mellon University, E-mails: zonggel@andrew.cmu.edu, hyunahs@cs.cmu.edu, christos@cs.cmu.edu
†University of Pittsburgh, E-mail: vladimir@sis.pitt.edu
‡University of Minnesota, E-mail: nikos@umn.edu



(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 1: **H-FUSE *is effective***: Tycho [13] (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*).

The advantages of our method are

- **Effectiveness**: our experimental result on real data show that H-FUSE reconstructs the original data with good accuracy, outperforming top competitors: (see figure 1).
- **Self-awareness** : H-FUSE provides an assessment of when the recovery is not reliable.
- **Scalability** : H-FUSE provably scales almost linearly on the length T of the target time sequence.

Figure 1 deserves some discussion: 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 (see Section 4.3 for more details).

Reproducibility: the Tycho dataset is publicly available [13]; our code is open-sourced at `https://github.com/zonggel/GFusion`.

The outline of the paper is typical: background, proposed method, experiments, theoretical analysis and conclusions.

2 Background and Related Work

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

2.1 Challenge of Historical Information Fusion

A major challenge in historical information fusion is estimating number of historical events from multiple aggre-

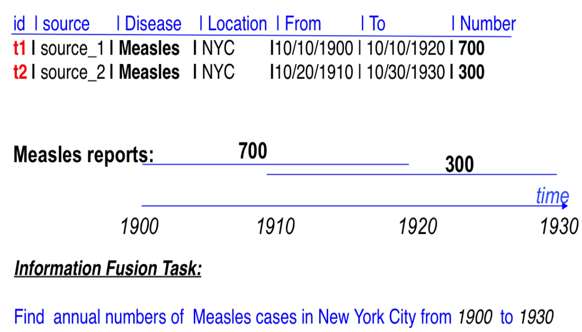
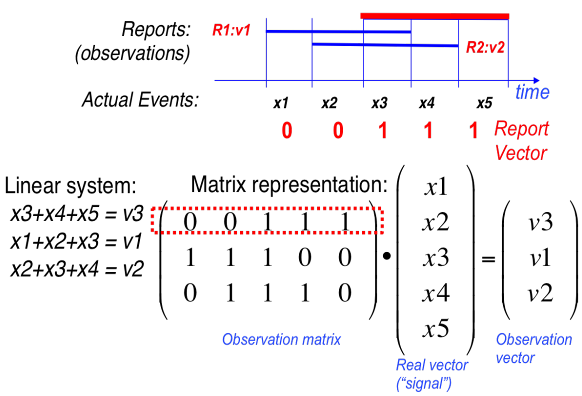


Figure 2: **Illustration on the nature of the data**. Time-overlapping historical reports

gated reports while handling redundant, and possibly, inconsistent information. Figure 2 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 disaggregation 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 3. 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.



$$\begin{pmatrix} 0 & 0 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 \end{pmatrix} \bullet \begin{pmatrix} x1 \\ x2 \\ x3 \\ x4 \\ x5 \end{pmatrix} = \begin{pmatrix} v3 \\ v1 \\ v2 \end{pmatrix}$$

Figure 3: **Illustration on our problem setting**. Characteristic Linear System of time-overlapping historical reports

2.2 Related Work Our work is related to a more general problem of a large-scale information integration to cope with diverse data sources. A prominent example of a large-scale information integration project is Tycho [13, 16, 17]. Currently, Tycho 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 in Section 4.

Historical information fusion task often requires dis-aggregation of historical reports and solving under-determined linear system. Temporal dis-aggregation methods have been studied in time series analysis of mostly economic data (see [2, 15, 3] for review). Given a low frequency time series (e.g. annual sales, weekly stock market index, etc.) the goal of temporal dis-aggregation is to produce a high-resolution series (e.g. quarterly sales, daily stock market index, etc.) while satisfying temporal aggregation constraints.

Temporal aggregation constraints ensure that the sum, average, the first or the last value of the resulting high frequency time series is consistent with the low frequency series. If available, related series observed at the required high frequency can be used to dis-aggregate the original observations. Such series are called indicators. However, care must be taken when selecting indicators since two strongly correlated low frequency time series may not be correlated at a higher frequency [6]. Thus, choosing good indicator series is not a straightforward task. 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.

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[4]*	*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 [18], vehicle detection [14], or super-resolution image reconstruction [11, 12]. A common way to find an approximate solution of an under-determined linear system is to apply least squares method (LSQ) and Tikhonov regularization [7, 5], by introducing additional constraints such as smoothness in space and/or temporal domain so that reconstruction reflects some aspects of the target data. Although Tikhonov regularization has been widely used in solving ill-posed problem in various communities, to our knowledge the application of Tikhonov regularization has not been addressed in historical data fusion domain.

In ’VLDB97’, Faloutsos et. al. [4] 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 (indicated as '?' in Table 1) but the method is completely incapable of handling conflicts.

3 Proposed 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

$$\mathcal{F}(\vec{x}) = \sum_{n=1}^{N}(v_n - \sum_{t=1}^{T} \mathbf{O}_{nt} x_t)^2 \tag{3.1}$$

and, in matrix form:

$$\mathcal{F}(\vec{x}) = ||\vec{v} - \mathbf{O}\vec{x}||_2^2 \tag{3.2}$$

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

3.1 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:

$$\min_{\vec{x}} \mathcal{L}(\vec{x}) = \min_{\vec{x}} \ (\mathcal{F}(\vec{x}) + \mathcal{C}(\vec{x})) \tag{3.3}$$

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. 3.2). 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:

$$\mathcal{C}_s(\vec{x}) = \sum_{t=1}^{T}(x_t - x_{t+1})^2 = ||\mathbf{H}_s \vec{x}||_2^2 \tag{3.4}$$

  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:

$$\mathcal{C}_p(\vec{x}) = \sum_{t=1}^{T}(x_t - x_{t+P})^2 = ||\mathbf{H}_p \vec{x}||_2^2 \tag{3.5}$$

  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 $t+P$.

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.

3.2 Methods 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 (Lemma 5.2).

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.

3.2.1 H-FUSE-S (Smoothness Method) The proposed loss function $\mathcal{L}_s(\cdot)$ is given by Equation 3.3

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

which gives, in matrix form:

$$\min_{\vec{x}} \mathcal{L}_s(\vec{x}) = \min_{\vec{x}} \ (||\vec{v} - \mathbf{O}\vec{x}||_2^2 + ||\mathbf{H}_s\vec{x}||_2^2) \tag{3.6}$$

3.2.2 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:

$$\min_{\vec{x}} ||\vec{v} - \mathbf{O}\vec{x}||_2^2 + 1/2||\mathbf{H}_s\vec{x}||_2^2 + 1/2||\mathbf{H}_p\vec{x}||_2^2 \tag{3.7}$$

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 3.1, page 4).

4 Experiments

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

4.1 Experimental setup To prove the effectiveness of H-FUSE on the real data, we apply H-FUSE on Tycho [13] 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.

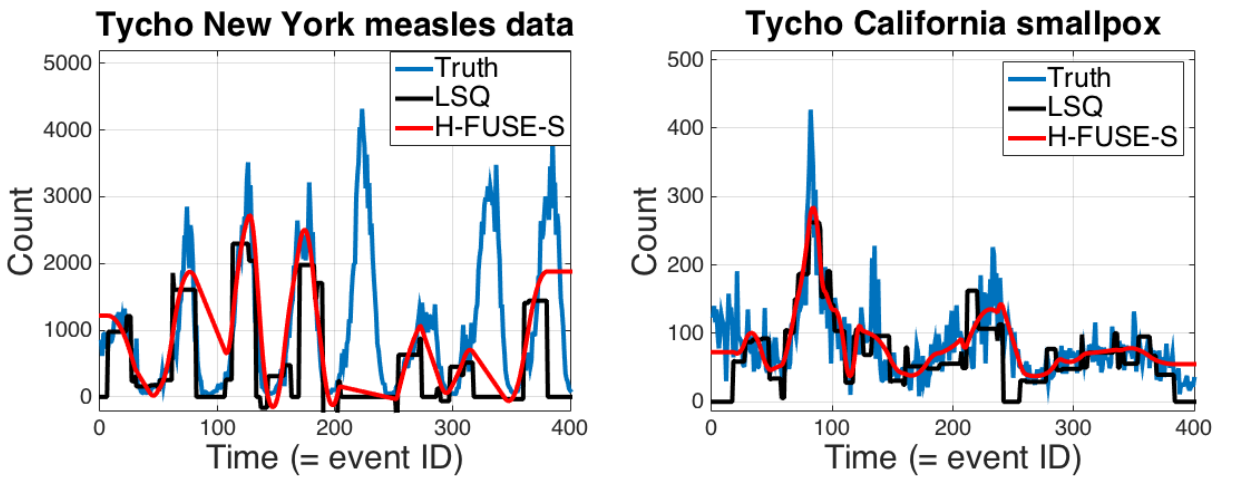
We refer to the observations or records as *reports*, and the number of observations as *report numbers*, and the timeticks that each report covers as *report duration*. We vary the report numbers and the report duration to conduct sensitivity test of H-FUSE. The reports are generated randomly for specific 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.

4.2 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 4, 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 5 (a). We vary the report number and the report duration in the x-axis 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.

4.3 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 3.2.2. Earlier in Figure 1, 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.

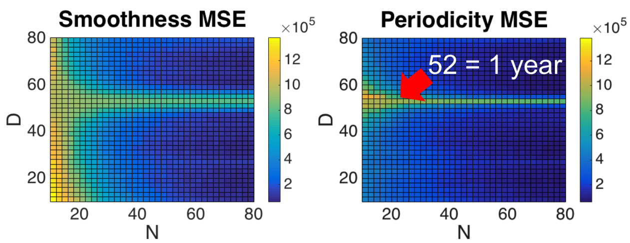
The error dynamics of H-FUSE-P on various settings of report number and report duration is shown in Figure 5 (b). We observe a similar trend as in the simple smoothness constraint case: the MSE decreases as the re-



(a) Tycho New York measles data - 30% improvement over LSQ

(b) Tycho California smallpox data - 58% improvement over LSQ

Figure 4: **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.



(a) H-FUSE-S

(b) H-FUSE-P

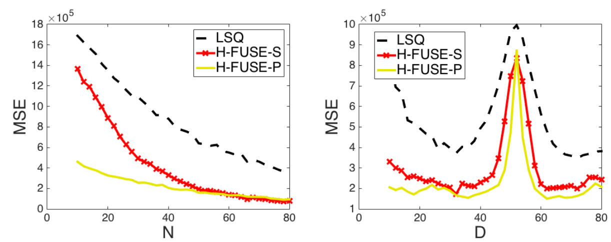
Figure 5: **H-FUSE with various configurations**. Both versions of H-FUSE reconstructs well (*blue*) unless the report duration matches exactly with the period $D = P (= 52)$.

port 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 5 (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 $N = 40$).

5 Theoretical analysis: Self-awareness, and Scalability

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.

5.1 Background Consider

$$\min_{\vec{x}} ||\vec{v} - \mathbf{O}\vec{x}||_2^2 + ||\mathbf{H}\vec{x}||_2^2 \tag{5.8}$$

which is equivalent to

$$\min_{\vec{x}} \left|\left| \begin{bmatrix} \vec{v} \\ - \\ \vec{0} \end{bmatrix} - \begin{bmatrix} \mathbf{O} \\ - \\ \mathbf{H} \end{bmatrix} \vec{x} \right|\right|_2^2 . \tag{5.9}$$

Here $\mathbf{H}$ is of type

$$\mathbf{H}_s = \begin{bmatrix} -1 & 1 & 0 & 0 & 0 & 0 \\ 0 & -1 & 1 & 0 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 & 0 \\ 0 & 0 & 0 & -1 & 1 & 0 \\ 0 & 0 & 0 & 0 & -1 & 1 \end{bmatrix}, \tag{5.10}$$

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.

5.2 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. 5.10.

LEMMA 5.1. *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 5.2. ■

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 5.2. *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*

$$SE = ||(\mathbf{O}^T\mathbf{O} + \mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{H}\vec{x}||_2^2 \tag{5.11}$$

Proof. Let $\hat{\vec{x}}$ be the solution we obtain from solving Equation 5.8. 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. 5.11 ■

Clearly, when $\mathbf{H}\vec{x} = 0$, the loss becomes zero, which is exactly what Lemma 5.1 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 5.3. *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 [4]. The proof is closely related to the Nyquist sampling frequency. ■

LEMMA 5.4. *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 5.3 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.

5.3 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 3.2.1. With the usual notation ($\vec{x}$ is the target sequence, of length T), we have the Lemma:

LEMMA 5.5. *For any report setting, let D_{max} be the longest report duration. Then the total computation time for our* H-FUSE-S *is*

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

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 [10]. ■

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

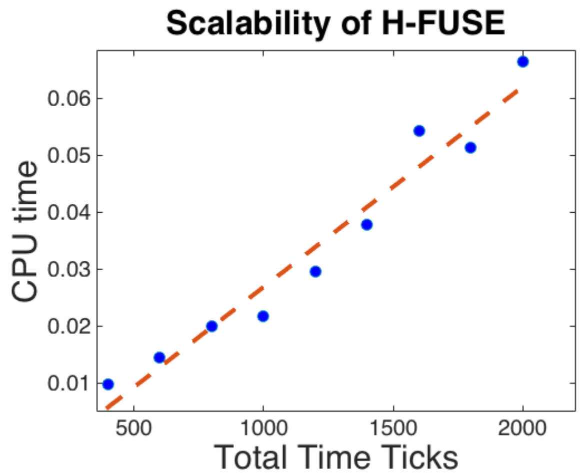


Figure 7: **H-FUSE is scalable** H-FUSE scales near-linearly on the length T. CPU time (seconds) vs T, in lin-lin scales; dashed orange line is plotted for reference.

"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 5.5 predicts.

6 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 5.2, and it is zero, if $\vec{x}$ is perfectly smooth (Lemma 5.1) 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 5.3, 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 5 gives us an arithmetic example where you have large MSE for D=52 weeks.

7 Conclusions

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. 3.3). 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 4.
2. **Self-awareness**: We provide theoretical results that help evaluate the quality of the reconstruction as discussed in Section 5.2.
3. **Scalability**: The computational cost for H-FUSE scales nicely as $O(Tlog(T) + 4D^2)$ as discussed in Section 5.3.

8 Acknowledgment

This material is based upon work supported 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. 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.

References

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

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

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

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

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

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

[7] G. H. Golub, P. C. Hansen, and D. P. O'Leary. Tikhonov regularization and total least squares. *SIAM Journal on Matrix Analysis and Applications*, 21(1):185–194, 1999.

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

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

[10] 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.

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

[12] V. Patanavijit and S. 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.

[13] T. Project:. https://www.tycho.pitt.edu.

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

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

[16] 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.

[17] 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.

[18] Y. Zhu, D. Comaniciu, M. Pellkofer, and T. Koehler. Reliable detection of overtaking vehicles using robust information fusion. *IEEE Transactions on Intelligent Transportation Systems*, 7(4):401–414, 2006.