Statistics and Application 统计学与应用, 2018, 7(2), 192-209
Published Online April 2018 in Hans. http://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2018.72024

Hans汉斯

Meteorological Data Restoration Based on Matrix Completion and Prior Features

Guanlei Xu[1,2*], Xiaotong Wang[1], Lijia Zhou[1], Limin Shao[1], Xiaogang Xu[2]

[1]Navigation Department of Dalian Navy Academy, Dalian Liaoning
[2]Ministry of Electronic Information and Engineering, Dalian University of Technology, Dalian Liaoning
Email: *xgl_86@163.com

Received: Apr. 1st, 2018; accepted: Apr. 21st, 2018; published: Apr. 28th, 2018

Abstract

Because of the limitation of observation means and background, combined with the complex environment, only some observation data are available. For the sake of better weather forecast, the research of meteorological data restoration based on part of observation data and matrix completion would have important scientific significance. This paper aims to, through part of real-time observation data, according to the low rank of a matrix, with applying SVT (Singular Value Thresholding) algorithm of matrix completion, obtain the deficient data so that one can make weather forecast better. The experimental result shows that the accuracy of forecast with matrix completion method is obviously higher than that with classical statistical method. When available data proportion is higher than the critical sampling proportion, errors of data filling can be controlled within 10%, which meet the requirements of meteorological data.

Keywords

Low Rank Matrix, Matrix Completion, SVT (Singular Value Thresholding) Algorithm, Prior Feature

基于矩阵优化填充和结构性先验统计信息的气象数据恢复

徐冠雷[1,2*]，王孝通[1]，周立佳[1]，邵利民[1]，徐晓刚[2]

[1]海军大连舰艇学院航海系，辽宁 大连
[2]大连理工大学电子信息与电气工程学部，辽宁 大连
Email: *xgl_86@163.com

收稿日期：2018年4月1日；录用日期：2018年4月21日；发布日期：2018年4月28日

*通讯作者。

文章引用: 徐冠雷, 王孝通, 周立佳, 邵利民, 徐晓刚. 基于矩阵优化填充和结构性先验统计信息的气象数据恢复[J]. 统计学与应用, 2018, 7(2): 192-209. DOI: 10.12677/sa.2018.72024

摘 要

由于观测手段和观测背景的限制，再加上环境复杂，很多时候只有部分气象观测资料可用，为了在这种背景下进行气象预报，充分完备的气象资料是重要基础，因此基于零散的部分观测数据、先验数据的统计特征和矩阵优化填充技术的气象资料恢复研究具有重要的工程价值和数学意义，其研究在国内外尚属空白。本文旨在通过部分观测资料，充分利用矩阵的低秩性和气象观测数据的内在结构性先验统计信息，应用矩阵填充的奇异值阈值化SVT算法，优化分析得到欠缺数据，从而获得填充的补全数据。实验结果表明，基于结构性先验信息和矩阵优化填充方法得到的数据准确率明显取决于矩阵格式选择和气象数据本身特性，而且本文通过理论和实验分析出最佳的矩阵优化填充模型，表明当可利用的资料占比高于临界采样率时，数据填充误差可控制在10%以内，可以有效地解决预报和分析时的观测资料数据缺失不全的问题。

关键词

低秩矩阵，矩阵填充，奇异值阈值化算法，先验统计特征

1. 引言

气象信息数据的获取是进行气象预报的前提，由于观测手段和观测背景的限制，再加上环境复杂，很多时候只有部分气象观测资料可用(可以获取)，为了在这种背景下进行气象预报，充分完备的气象资料是重要基础，因此基于部分观测数据和矩阵填充的气象资料恢复研究具有重要的科学意义[1]。本文旨在通过部分观测资料，根据矩阵的低秩性和气象观测数据的内在结构性先验信息，应用矩阵填充的 SVT 算法，得到欠缺的数据，从而获得补全数据。

近几年来，在工程控制、机器学习和计算机视觉等应用科学的众多领域，人们越来越感兴趣于如何利用非常有限的信息恢复出满足某种条件的未知信号[2]，这就涉及到如何对稀疏信号进行有效编码的问题[3]。例如，文献[4]提出的压缩感知(compressed sensing，CS)理论即通过少数的随机测量值完全或近似恢复出未知信号。与压缩感知理论相类似，矩阵填充(matrix completion，MC)理论也是通过对未知信号进行有效的降维处理，然后求解相关的稀疏信号重构问题，即通过少量的稀疏采样值完全或近似恢复出原始信号。矩阵填充的一个典型例子是 Netflix 问题[5]，该公司让用户在观看影碟后对电影打分，然后该公司根据用户对于影片的打分数据对用户喜好进行预测，设计新的预测方法。矩阵填充理论已被广泛的运用到量子理论、人脸识别、在线推荐系统等很多实际问题中，当然也包括气象领域的问题。比如，对于海上航行的船舶，由于观测手段和观测背景的限制，再加上环境复杂，通信往往出现断断续续甚至中断的情况，只有部分随船观测资料和历史背景资料可用，为了在这种背景下进行水文气象预报保障服务，基于部分观测数据和矩阵填充的天气预报的研究具有重要的科学意义和军事意义。还有，对于无人值守的海岛气象自动观测站，由于系统可靠性等原因，观测数据丢失是会经常发生的，面对该问题，目前除临近插值(或同化)外，别无它法。

同时，研究经验表明，气象数据具有很强的结构性和周期性(先验特征信息) [1] [6]，在矩阵格式下就

表现为低秩性，本文将结合气象数据矩阵的低秩性，应用矩阵填充技术 SVT 进行缺失数据的补全。

2. 矩阵填充

一个矩阵如果只观测到了它的一部分元素(这一部分元素可能占该矩阵元素很低的比例)，我们要推测出它的其他没有观测到的元素的信息，这便是矩阵填充所研究的问题。如果对于矩阵没有任何条件限制，矩阵填充问题的解将是无穷多的，是不可解的[5] [7]，但是在实际问题中，很多时候我们遇到的矩阵都是低秩矩阵或者渐进低秩矩阵，比如本文利用到的历史资料矩阵便是低秩矩阵，对于低秩矩阵，研究表明，可以通过合适的方法准确恢复出原来的矩阵。矩阵填充问题总结起来就是求解核范数最小化问题[7]。虽然现在求解核范数最小化问题已经有了一些成熟的算法，但是目前这些算法的复杂度都很高，处理高维矩阵的填充问题会花费大量时间，也正因此，快速高效的矩阵填充算法也是矩阵填充问题的一个研究热点。目前已经诞生出很多快速高效的矩阵填充算法[8]-[14]，最早的是由 Cai 等人提出的 SVT 算法[8]，该算法受到压缩感知中 Bregman 迭代算法的启发，算法在迭代过程中对矩阵进行了奇异值分解，然后将较小的奇异值设置为 0，生成新的矩阵进行反复迭代，此算法运行速度很快，对于高维低秩矩阵的回填效果很好。另外还有 Ma 等人提出的 FPCA 算法[9]，该算法也用到了矩阵的奇异值分解，并且在算法迭代过程中进行不动点连续处理。该算法不仅对于低秩矩阵的恢复效果很好，对于秩适中的矩阵也有较好的恢复效果。在这之后，又陆续出现了很多关于矩阵填充的快速高效算法。目前矩阵填充问题与算法的研究已经取得了极大的进展，但是理论的不成熟导致仍然存在很多问题，例如一些实际问题中需要填充的低秩矩阵，其核范数是固定的，此时应用核范数最小化方法求解显然没有意义，对于这些问题，需要提出新的算法。另外，矩阵填充理论在各领域的应用也是一个重要的研究方向，特别是气象领域。由于 Cai 等人提出的 SVT 算法[8]是目前最为流行的算法，也是应用最为广泛的算法，本文将采用该算法进行气象数据的填充。SVT 算法基本思路如下：

对于给定的矩阵 X，矩阵部分数据已知，即下面优化问题即是矩阵填充的数学模型：

$$\min \quad \|X\|_*, \text{ s.t. } X_{i,j} = M_{i,j}, \ (i,j) \in \Omega, \tag{1}$$

如果矩阵中数据采样对于给定的某个常数 C 满足 $m \geq Cn^{6/5} r \log n$，上式就会以较高的概率$(1-n^{-3})$恢复出矩阵缺失元素。这里$\|X\|_*$表示的是矩阵的核范数，即所有奇异值的和，r 为矩阵的秩，n 为矩阵行数和列数中的最小值，m 为矩阵中已知数据个数。由于求解(1)较为困难，上式可以松弛成如下优化问题：

$$\min \quad rank(X), \text{ s.t. } X_{i,j} = M_{i,j}, \ (i,j) \in \Omega. \tag{2}$$

进一步，Cai 等人把限制条件进行了改进[8]，不是直接相等约束，而是投影(投影算子设为 P_Ω)后具有相同的数值(即改为投影约束)，即：

$$\min \quad rank(X), \text{ s.t. } P_\Omega(X_{i,j}) = P_\Omega(M_{i,j}), \ (i,j) \in \Omega. \tag{3}$$

因此，可以通过迭代优化计算方法(4)直到达到某个停止条件，获得最终的优化矩阵 X：

$$\begin{cases} X^k = shrink\left(Y^{k-1}, \tau\right) \\ Y^k = Y^{k-1} + \delta_k P_\Omega\left(M - X^k\right) \end{cases}, \tag{4}$$

其中，$Y^0 = 0$，$shrink\left(Y^{k-1}, \tau\right)$为一个非线性软阈值函数，阈值为$\tau$，$\delta_k$ 为 k 步相应的步长。这里使用了两个重要的特性：稀疏性和低秩性。矩阵在迭代的过程中一直保持着稀疏性，同时矩阵必须是低秩的，否则该方法将失效。

很显然，矩阵填充问题是一个非适定性的问题。一般而言，如果一个矩阵仅仅由少量的采样元素组

成，那么完全重构出原矩阵几乎是不可能的，因为对矩阵未知元素的填充有无限种可能性。如果没有其他约束条件，矩阵填充重构出的矩阵将不是唯一的。但是如果我们事先知道原始矩阵数据满足一定的条件，那么矩阵填充将变得可行，这个关键的条件就是矩阵的低秩性[10]。

在实际的气象数据处理问题中，我们希望恢复的未知矩阵往往都是低秩的或近似低秩的。文献[12]证明了未知矩阵的低秩性或近似低秩性是矩阵填充重构出矩阵唯一性的前提，未知矩阵的低秩性或近似低秩性从根本上改变了矩阵填充问题的非适定性。该文还证明了通过解决一个简单的凸优化问题，就可以用极其少量已知的采样元素精确填充得到未知的低秩矩阵[13]。

本文主要利用的就是气象观测数据矩阵的低秩性或近似低秩性来填补构造矩阵。在实际情况中，由于观测条件和人为等因素，气象数据有时候会出现缺失(甚至缺失的比例较大)，面临很大程度的不完整性，而如何利用已有的部分观测数据获取较为准确且结构性良好的整体数据，这便是本文要解决的问题。

3. 气象数据的先验特征

本文选取了大连海区 2011 年至 2015 年 5 年共计 1825 天完整历史天气数据，数据下载网站为 http://tianqi.2345.com/wea_history/54662.htm。其中每一天的数据包括日期、最高气温、最低气温、天气类型、风向以及风力等 5 类数据。部分数据如表 1，表中数据完整，数据量较大，作为本文研究的数据。另外，还可以利用经典统计方法对 5 年的数据进行相应的统计，获取经验性成果，提高天气形势反演的准确率并验证之。

Table 1. Historical Weather Data of Dalian (Part)
表 1. 大连历史天气数据(部分)

日期	最高气温	最低气温	天气	风向	风力
2011-01-02 星期日	−3 ℃	−7℃	多云~晴	北风~西北风	4~5 级
2011-01-03 星期一	−3℃	−8℃	晴	西北风~北风	4~5 级
2011-01-04 星期二	−3℃	−8℃	多云	西北风~北风	5~6 级
2011-01-05 星期三	−4℃	−6℃	多云~晴	北风~西北风	5~6 级
2011-01-06 星期四	−3℃	−6℃	晴	西北风~北风	4~5 级
2011-01-07 星期五	−5℃	−8℃	晴	北风	4~5 级
2011-01-08 星期六	−5℃	−6℃	晴	北风	4~5 级
2011-01-09 星期日	−1℃	−7℃	晴	西南风	4~5 级
⋮	⋮	⋮	⋮	⋮	⋮
2015-12-25 星期五	5℃	−3℃	晴	西南风~北风	5~6 级
2015-12-26 星期六	−4℃	−10℃	晴	北风	6~7 级
2015-12-27 星期日	−5℃	−12℃	晴~多云	北风~西北风	6~7 级~4~5 级
2015-12-28 星期一	2℃	−2℃	晴~多云	西北风~西南风	5~6 级~4~5 级
2015-12-29 星期二	4℃	0℃	多云~阴	西南风~北风	4~5 级
2015-12-30 星期三	2℃	−4℃	阴~多云	西北风	5~6 级
2015-12-31 星期四	2℃	−3℃	晴	西北风~西南风	4~5 级

3.1. 数据的规律统计特性

由于各物理量之间存在着明显的规律性，比如当时间为冬季的时候，最低气温将会接近其极小值，天气很可能会出现下雪天气；当风力为大风时基本不可能出现大雾天气等等。加之数据量较大，对数据进行统计得到的结果对于本文研究的先验特征具有重要支持依据。

由图 1 可以看出，自 2011 年 01 月 01 日到 2015 年 04 月 14 日，大连海区共出现晴 613 天，多云 466 天，阴 41 天，雾 36 天，雨 306 天，雪 86 天，沙尘 4 天。其中最常见的天气是晴天，共出现 613 天，约占 39.5%；沙尘天气最少，仅仅出现过 4 天，不足 1.0%，在后续的数据处理上，由于其占比太低，故将沙尘天气略去；多云天气以及雨雪天气也较为常见，比例分别为 30.0%、25.3%，阴、雾天气较少，比例约为 2.6%、2.3%。

由图 2 可以看出，自 2011 年 01 月 01 日到 2015 年 04 月 14 日，大连海区共出现北风 390 天，东北风 142 天，东风 52 天，东南风 196 天，南风 249 天，西南风 191 天，西风 66 天，西北风 266 天。大连海区主要以偏北风和偏南风为主，其中仅正北风和正南风而言，二者就占了约 41.2%；而东风和西风相对而言出现的频率非常低，仅占约 7.6%。这个规律性相当明显，在后续的数据填充中应当加以考虑。

由图 3 可以看出，自 2011 年 01 月 01 日到 2015 年 04 月 14 日，大连海区共出现 3-4 级风 5 天，4-5 级风 913 天，5~6 级风 525 天，6~7 级风 104 天，7~9 级风 5 天。4~5 级风是大连地区最常见的风，占了 58.8%，频率在一半以上；5~6 级风也比较常见，约占 33.8%；4~5 级风、5~6 级风加起来共占 92.7%，说明大连地区风的风力基本保持在 4~6 级左右，其他风力较为少见。由图 4 可以看出，自 2011 年 01 月 01 日到 2015 年 04 月 14 日，大连海区共出现天最高气温−11~−6 ℃ 21 天，天最高气温−5~−1 ℃ 151 天，天最高气温 0℃~5℃共 281 天，天最高气温 6℃~10℃共 168 天，天最高气温 11~15℃共 163 天，天最高气温 16℃~20℃共 183 天，天最高气温 21℃~25℃共 295 天，天最高气温 26℃~32℃共 290 天。

由图 5 可以看出，自 2011 年 01 月 01 日到 2015 年 04 月 14 日，大连海区共出现天最低气温−15℃~−11℃共 24 天，天最低气温−10℃~−6℃共 174 天，天最低气温−5℃~−1℃共 247 天，天最低气温 0℃~5℃共 237 天，天最低气温 6℃~10℃共 172 天，天最低气温 11℃~15℃共 211 天，天最低气温 16℃~20℃共 288 天，天最低气温 21℃~25℃共 199 天。

由图 6 可以看出，自 2011 年 01 月 01 日到 2015 年 04 月 14 日，大连海区共出现温差 0℃~3℃共 82 天，温差 4℃~7℃共 1028 天，温差 8℃~11℃共 420 天，温差 12℃~15℃共 22 天。大连地区一天的温差

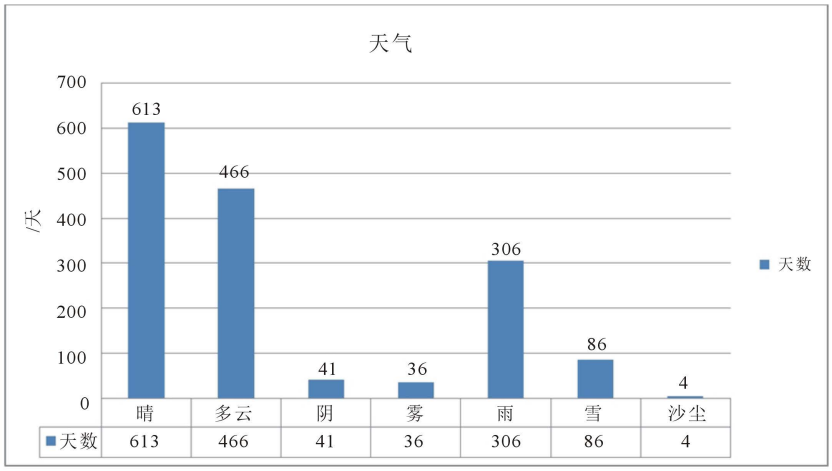

Figure 1. The historical statistical weather class data of Dalian
图 1. 大连历史天气现象统计

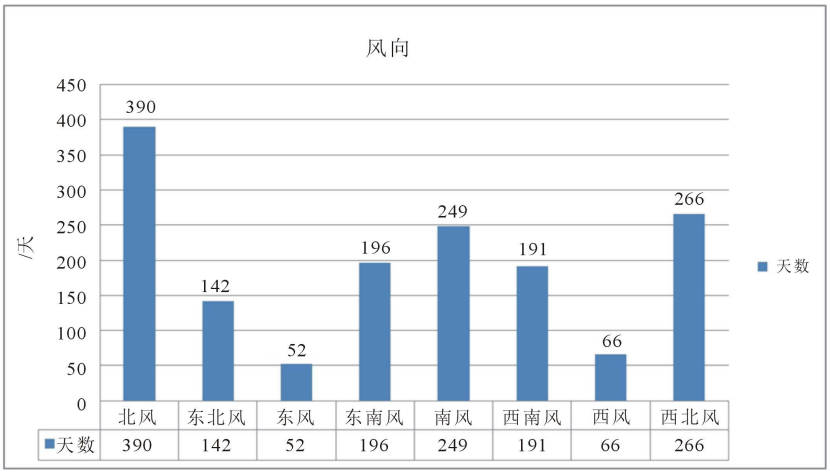

Figure 2. The historical statistical wind direction data of Dalian
图 2. 大连历史风向数据统计

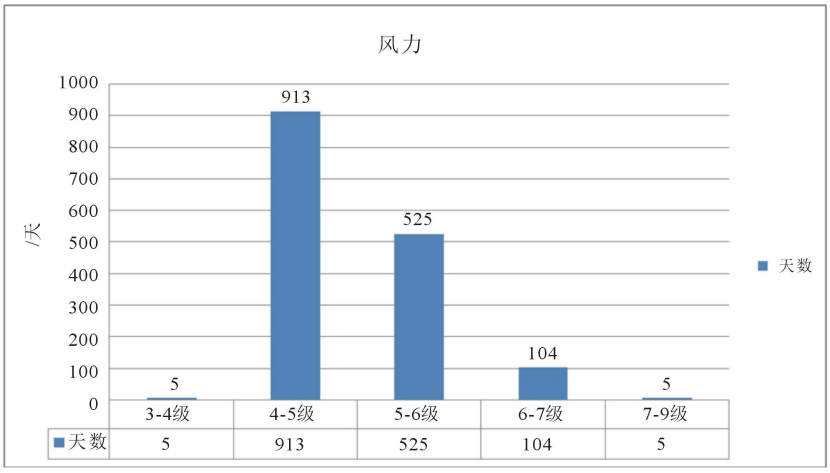

Figure 3. The historical statistical wind power data of Dalian
图 3. 大连历史风力数据统计

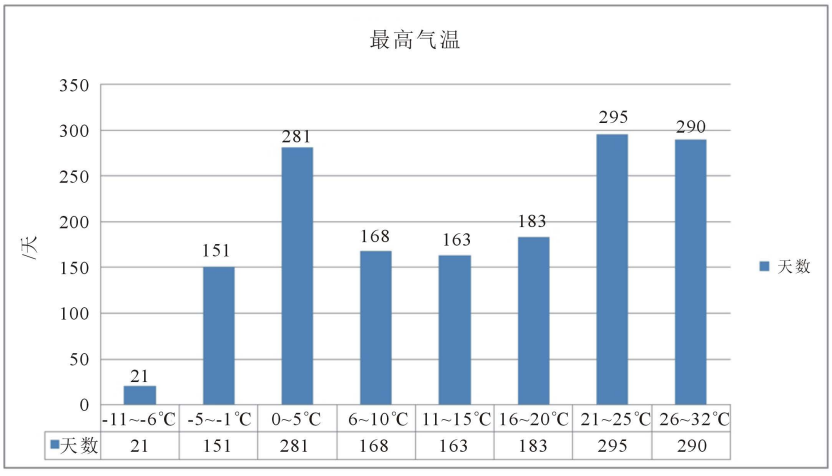

Figure 4. The historical statistical highest temperature data of Dalian
图 4. 大连历史最高气温资料统计

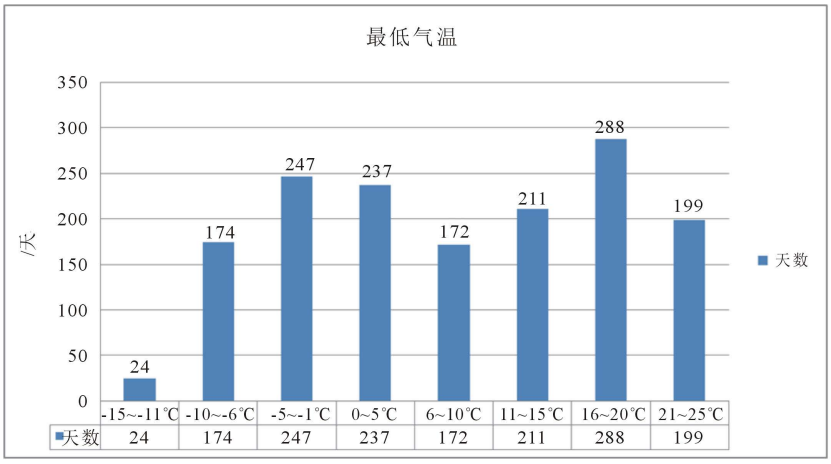

Figure 5. The historical statistical lowest temperature data of Dalian
图 5. 大连历史最低气温资料统计

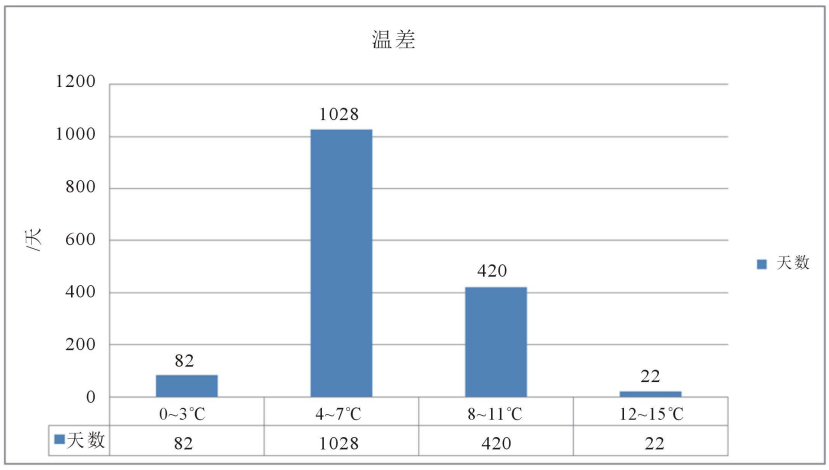

Figure 6. The historical statistical temperature difference data of Dalian
图 6. 大连历史温差资料统计

主要在 4℃~7℃左右，说明大连海区一天的温度变化并不大，这与大连海区的地理位置有很大的关系：大连地区靠海，海水的比热容比陆地的大，导致大连地区气温变化不剧烈。

正是上述数据的这种规律统计特性(图 1~6)，可以说为数据缺失的填充恢复提供了物理基础。事实证明，这种数据的内在统计规律预示了矩阵内在的结构性和稀疏性，为数据填充可行性提供了依据。

3.2. 数据的分类

在日常生活中我们会发现，不同季节、不同天气的温差是不一样的，这说明温差与它们可能存在一定的关系，如果确实存在这样的关系，那么在矩阵中将会得到体现，即矩阵的低秩性特征。故在此将最高气温减去最低气温得到每日温差数据，以扩展我们的数据矩阵，得到更好的填充效果。经初步处理后部分数据如表 2 所示。

我们最终要将大连历史天气数据转化成矩阵来使用，而原始数据是由文本组成的，且种类繁多，这需要我们对原始数据进行分类，将相似的物理现象归为一类。下面我们就根据气象数据本身的特征进行分类。

Table 2. Historical weather data of Dalan (Part)
表 2. 大连历史天气数据(部分)

日期	最高气温	最低气温	温差	天气	风向	风力
2011-01-01 星期六	−4℃	−8℃	4℃	晴~多云	北风	5-6 级~4-5 级
2011-01-02 星期日	−3℃	−7℃	4℃	多云~晴	北风~西北风	4~5 级
2011-01-03 星期一	−3℃	−8℃	5℃	晴	西北风~北风	4~5 级
2011-01-04 星期二	−3℃	−8℃	5℃	多云	西北风~北风	5~6 级
2011-01-05 星期三	−4℃	−6℃	2℃	多云~晴	北风~西北风	5~6 级
2011-01-06 星期四	−3℃	−6℃	3℃	晴	西北风~北风	4~5 级
2011-01-07 星期五	−5℃	−8℃	3℃	晴	北风	4~5 级

1) 天气类型分类

由于大连海区这 5 年的数据中天气类型的种类十分复杂，共计 123 种，而人们日常关心更多的是恶劣天气，所以如果某天的天气类型没有变化，就取当日天气作为最终天气，例如 2011 年 01 月 03 号天气为晴，则最终天气类型便为晴天；如果某天的天气类型有所改变，则取恶劣天气做为当日天气类型，以此简化数据种类，例如 2011 年 04 月 18 日天气为阵雨~晴，则取“阵雨”为当日分析天气。另外，为了进一步简化数据，削减数据种类，在处理数据时将含有雪的天气都归为一类“雪”，例如大雪、中雪、小雪最后都定为“雪”，按照此方法，最终天气类型被简化成 6 种基本天气类型，分别是雪、雨、雾、阴、多云、晴。

2) 风向分类

风向和天气一样，种类多样，比较复杂，共计 64 种。我们研究风矢量的时候和研究天气一样，更多的是关心大风的影响(而非常细化的风向考虑相对少一些)。故如果某日风向无变化则取该风向为当日风向；若风向改变，但风力不变，则取初始风向为当日风向；如果风向改变并且风力也改变，则取较大风力等级对应的风向作为当日风向。例如 2011 年 01 月 01 日风向为北风，则当日风向为北风；2011 年 01 月 02 日风向为北风－西北风，风力 4~5 级未变，则取初始风向北风作为当日风向；而 2011 年 01 月 21 日风向为东北风－北风，风力为 6~7 级~5~6 级，则取较大风力 6~7 级对应风向东北风作为当日风向。最终风向简化成 8 类，分别是北风、东北风、东风、东南风、南风、西南风、西风、西北风。

3) 风力分类

日常生活中由于我们更多地关心大风的风力影响，所以对于风力的分类处理比较简单，如某日风力未变则该风力为当日风力；如果风力改变则取大风力为当日风力。例如 2011 年 01 月 19 日风力为 4~5 级~5~6 级，则当日风力为 5~6 级。

4)气温温差分类

将天最低气温、天最高气温和日气温差数据中的摄氏度符号“℃”直接去掉，然后作为其等级即可。经分类处理之后部分数据如表 3 所示。

3.3. 数据的结构性等级划分

经过上面的数据分类处理之后，我们在合理范围内将数据的种类降低，这对于后续的矩阵处理是非常有利的，而且因为我们更多是关心恶劣天气的影响，所以这对最终数绝填充结果影响并不大。现在

Table 3. Historical weather data after classification (Part)
表 3. 经分类的大连历史天气数据(部分)

日期	最高气温	最低气温	温差	天气	风向	风力
2011-01-01 星期六	−4	−8	4	多云	北风	5~6 级
2011-01-02 星期日	−3	−7	4	多云	北风	4~5 级
2011-01-03 星期一	−3	−8	5	晴	西北风	4~5 级
2011-01-04 星期二	−3	−8	5	多云	西北风	5~6 级
2011-01-05 星期三	−4	−6	2	多云	北风	5~6 级
2011-01-06 星期四	−3	−6	3	晴	西北风	4~5 级
2011-01-07 星期五	−5	−8	3	晴	北风	4~5 级
2011-01-08 星期六	−5	−6	1	晴	北风	4~5 级
2011-01-09 星期日	−1	−7	6	晴	西南风	4~5 级
2011-01-10 星期一	−4	−7	3	阴	北风	5~6 级
2011-01-11 星期二	−3	−8	5	晴	北风	4~5 级
2011-01-12 星期三	−4	−7	3	晴	西南风	4~5 级
2011-01-13 星期四	−4	−9	5	晴	北风	4~5 级

我们要将数据进行相应的编号，即结构性等级划分的数字化，得到新的数字矩阵。这样便可以将大连海区的历史数据转化成数据矩阵处理(表 4~9)。

经数字化处理之后部分数据如表 10 所示(日期数据未处理)：

时间维对于整个大连海区历史天气数据矩阵而言是非常重要的一维，因为时间对应的一年四季，会出现哪样的天气现象是十分有规律性的，它包含着丰富的物理意义，与其它物理量的内在联系也非常明显，故在此将时间维分四类进行数字化处理，并在之后的处理中比较各类处理方法的优缺点。

1) 按月处理(表 11)

1.1) 竖排结构

将 5 年 1825 天的数据排成 1825 × 7 的矩阵，顺序不变。

1.2) 横排结构

将 5 年 1825 天的数据排成 365 × 35 的矩阵，其中 1~7 列为 2011 年的数据，8~14 列为 2012 年的数据，15~21 列为 2013 年的数据，22~28 列为 2014 年的数据，29~35 列为 2015 年的数据。

2) 按季处理(表 12)

2.1) 竖排结构

和按月处理竖排一样，将 5 年 1825 天的数据排成 1825 × 7 的矩阵，顺序不变。

2.2) 横排结构

同样，将 5 年 1825 天的数据排成 365 × 35 的矩阵，其中 1~7 列为 2011 年的数据，8~14 列为 2012 年的数据，15~21 列为 2013 年的数据，22~28 列为 2014 年的数据，29~35 列为 2015 年的数据。

经过将数据进行数字化处理，最终得到了按月处理(竖排)、按月处理(横排)、按季处理(竖排)、按季处理(横排)四种类型的数据。之所以这么做是因为不同结构的数据各物理量表现内在联系的方式不一样，所得到的矩阵的秩也是不一样的，运用这四种结构分别处理得到结果，再进行比较得到最优结构。

Table 4. Digitization of highest temperature
表 4. 最高气温数字化

范围	−15℃~−6℃	−5℃~−1℃	0℃~5℃	6℃~10℃
编号	1	2	3	4
范围	11℃~15℃	16℃~20℃	21℃~25℃	26℃~35℃
编号	5	6	7	8

Table 5. Digitization of Lowest Temperature
表 5. 最低气温数字化

范围	−15℃~−11℃	−10℃~−6℃	−5℃~−1℃	0℃~5℃
编号	1	2	3	4
范围	6℃~10℃	11℃~15℃	16℃~20℃	21℃~25℃
编号	5	6	7	8

Table 6. Digitization of Temperature Difference
表 6. 温差数字化

范围	0℃~4℃	5℃~7℃	8℃~10℃	11℃~15℃
编号	1	2	3	4

Table 7. Digitization of Weather Class
表 7. 天气类型数字化

天气类型	晴	多云	阴	雾	雨	雪
编号	1	2	3	4	5	6

Table 8. Digitization of Wind Direction
表 8. 风向数字化

风向	北风	东北风	东风	东南风	南风	西南风	西风	西北风
编号	1	2	3	4	5	6	7	8

Table 9. Digitization of Wind Power
表 9. 风力数字化

风力	3~5 级	5~6 级	6~9 级
编号	1	2	3

4. 基于部分数据和先验特征的气象数据填充

本文在研究前人的系列工作上，首先将已经数字化的大连历史天气数据输入到不同矩阵格式中(按季节、按月、竖排、横排等)，再利用矩阵填充的 SVT 算法对数据进行填充。为了测试算法的效能，我们利用不同的采样率随机采样以得到各采样率下的矩阵填充相对误差，比较得到多高的采样率下矩阵填充相对误差能控制在 10.00%以内。运算结果如表 13 所示。

Table 10. Digitization of historical weather data of Dalian (Part)
表 10. 经数字化的大连历史天气数据(部分)

日期	最高气温	最低气温	温差	天气	风向	风力
2011-01-01 星期六	2	2	1	2	1	2
2011-01-02 星期日	2	2	1	2	1	1
2011-01-03 星期一	2	2	2	1	8	1
2011-01-04 星期二	2	2	2	2	8	2
2011-01-05 星期三	2	2	1	2	1	2
2011-01-06 星期四	2	2	1	1	8	1
2011-01-07 星期五	2	2	1	1	1	1
2011-01-08 星期六	2	2	1	1	1	1

Table 11. Processing the data according to months
表 11. 按月处理数字化

月份	1 月份	2 月份	3 月份	4 月份	5 月份	6 月份
编号	1	2	3	4	5	6
月份	7 月份	8 月份	9 月份	10 月份	11 月份	12 月份
编号	7	8	9	10	11	12

Table 12. Processing the data according to seasons
表 12. 按季处理数字化

季节	春季	夏季	秋季	冬季
月份	3、4、5 月	6、7、8 月	9、10、11 月	12、1、2 月
编号	1	2	3	4

在相对填充误差 10%左右的范围在进一步细化采样率，得到相应的误差率，从而找出相对填充误差为 10%对应的临界采样率，结果如下(表 14)：

将表 15 中的数据细化(采取更小的步长)分别做成曲线图，如图 7、图 8 所示。

由相对填充误差图可以看出，随着数据采样率的提高，误差在不断减小。其中当数据采样率在 50%以下时，随着数据采样率的提高，误差在迅速减小；而当数据采样率在 50%以上时，随着数据采样率的提高，误差减小的速率较慢。另外由图可以看出，无论数据是按月处理还是按季处理，横排结构的数据误差都要小于竖排结构的误差，即横排结构优于竖排结构；而观察按月处理或是按季处理，可以看出在正常范围内无论是横排结构还是竖排结构，数据按月处理的误差均小于按季处理的误差，即按月处理优于按季处理。所以最优数据结构为按月处理横排结构，此时数据填充的误差最小，天气数据填充的效果最好。要将误差控制在 10.00%以内的的临界采样率如表 16 所示。

5. 按月横排矩阵格式最优化分析

根据 SVT 算法中矩阵低秩的特性，即对于同样的数据秩最低时 SVT 填充效果最佳，根据优化方程式(1)

Table 13. The error rate under different sampling rate
表 13. 不同采样率下的误差率

采样率	按月处理(竖排)	按月处理(横排)	按季处理(竖排)	按季处理(横排)
	相对填充误差	相对填充误差	相对填充误差	相对填充误差
5%	97.37%	92.12%	97.43%	94.27%
10%	75.46%	70.62%	77.81%	79.56%
15%	70.03%	62.29%	76.68%	46.74%
20%	52.53%	28.38%	66.44%	33.09%
25%	43.39%	11.35%	52.27%	21.06%
30%	34.77%	8.17%	48.22%	13.31%
35%	24.63%	7.28%	30.46%	10.67%
40%	19.29%	5.38%	27.12%	8.66%
45%	13.73%	5.37%	20.35%	6.74%
50%	11.71%	4.34%	16.94%	6.22%
55%	9.56%	3.63%	13.52%	5.42%
60%	7.64%	3.17%	11.99%	4.53%
65%	6.21%	3.03%	8.87%	3.78%
70%	4.44%	2.33%	7.35%	3.64%
75%	3.42%	1.99%	5.86%	2.65%
80%	2.43%	1.53%	4.03%	2.27%
85%	1.66%	1.05%	2.56%	1.46%
90%	1.00%	0.77%	1.71%	0.99%
95%	0.55%	0.37%	0.82%	0.53%
100%	0.00%	0.00%	0.00%	0.00%

Table 14. The error rate according to months
表 14. 按月处理误差率

采样率	按月处理(竖排)	采样率	按月处理(横排)
50%	12.15%	25%	11.13%
51%	11.79%	26%	10.56%
52%	10.67%	**27%**	10.22%
53%	10.26%	**28%**	9.68%
54%	9.99%	29%	8.95%
55%	9.79%	30%	8.78%

可知，如果矩阵中数据采样对于给定的某个常数 C 满足 $m \geq Cn^{6/5} r \log n$，上式就会以较高的概率$(1-n^{-3})$恢复出矩阵缺失元素。所以我们不妨定义一个矩阵恢复测度 ξ：

Table 15. The error rate according to seasons
表 15. 按季处理误差率

采样率	按季处理(竖排)	采样率	按季处理(横排)
60%	11.54%	30%	13.81%
61%	11.09%	31%	11.46%
62%	10.51%	32%	11.24%
63%	10.25%	33%	11.21%
64%	9.41%	**34%**	10.88%
65%	9.35%	**35%**	9.70%

Table 16. The threshold sampling rate
表 16. 临界采样率

数据结构	按月处理(竖排)	按月处理(横排)	按季处理(竖排)	按季处理(横排)
临界采样率	53.7%	27.3%	63.2%	34.7%

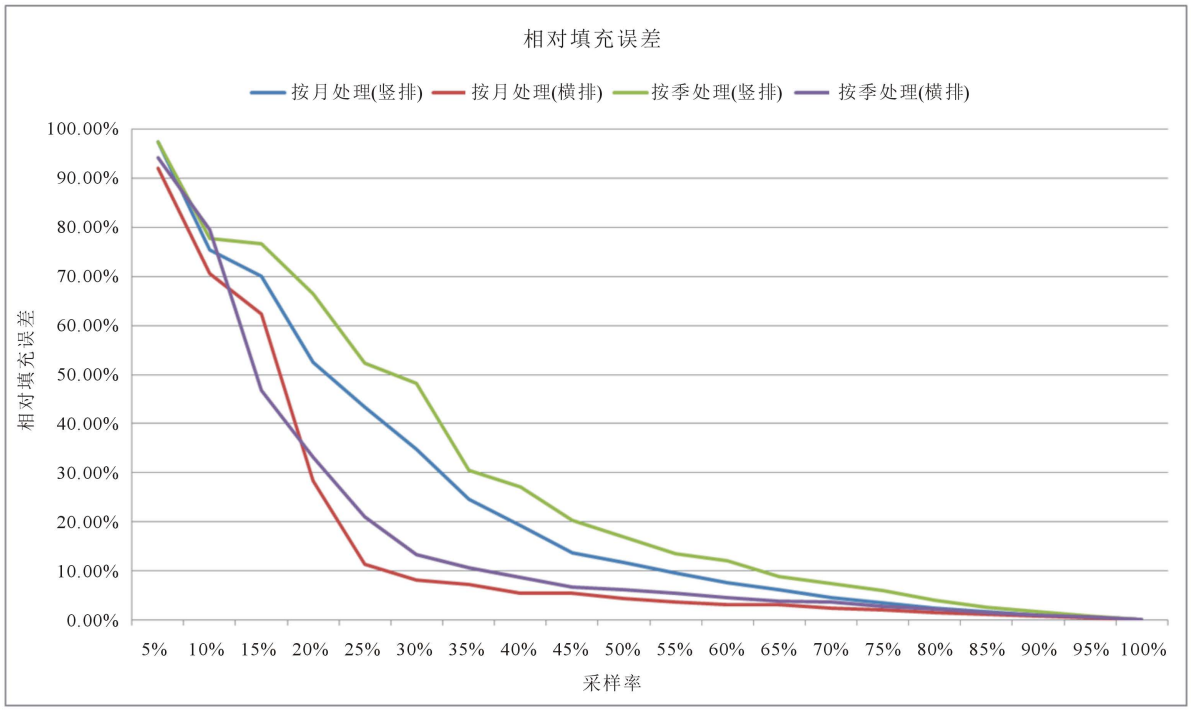

Figure 7. The error rate under different sampling rate
图 7. 不同采样率下的误差率

$$\xi \propto \frac{M}{m} \cdot \frac{r}{n}, \tag{5}$$

其中，r 为矩阵的秩且满足 $r \le n$，n 为矩阵行数和列数中的最小值，m 为矩阵中已知数据个数，M 为矩阵数据总个数即总行数和总列数的乘积。显然，ξ 越小，表明矩阵越易于恢复填充，相反 ξ 越大，表明

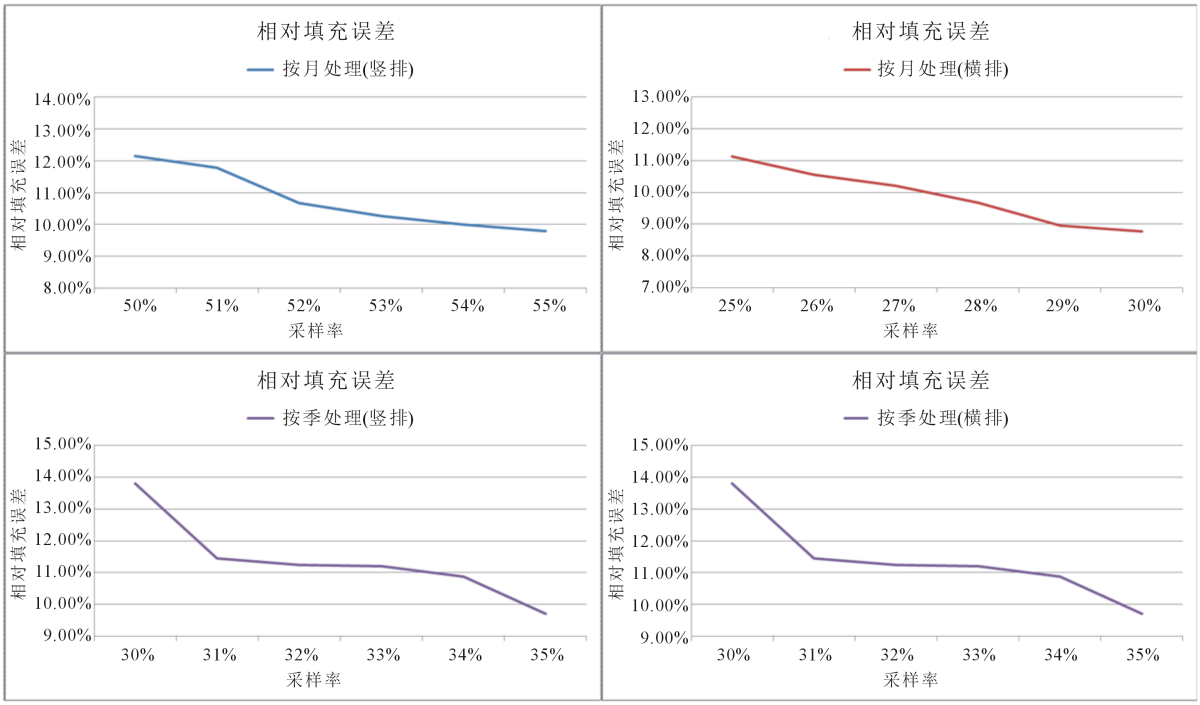

Figure 8. The error rate under different sampling rate (refined)
图 8. 不同采样率下的误差率(细化)

矩阵越难于恢复填充。这一点很容易从直观上去解释，如果已知的数据个数 m 越大(其他参数固定)，显然矩阵需要填充的数据就越少，越容易填充，极限情况则是没有数据缺失 $m=M$，填充率可达 100%。相反，如果 $m=0$ 即没有已知数据，则 $\xi\to+\infty$ 表示矩阵无法填充。另一方面，如果已知的数据个数 m 已定的情况下，矩阵的秩 r 越大，ξ 越大，表明矩阵越难于恢复填充，如果 $r=n$，表示满秩，数据填充困难(如果有数据缺失的话)。相反，矩阵的秩 r 越小，ξ 越小，表明矩阵越易于恢复填充，如果 $r=1$，且 $n\gg1$ 时，$\xi\to0$，表明矩阵数据中只有一行或列就可以表出其他行列数据。特殊情况，如果 $r=0$，表示矩阵为 0 矩阵，直接不必填充即可获得结果。

所以，我们只需证明按月横排数据时的矩阵具有最小的矩阵恢复测度 ξ 即可。

假定有 T 年数据，每年固定数据量为 p 个，每个数据包含 q 个变量(比如日期，最高气温，最低气温，温差，天气，风向，风力等 q 个要素)，这些数据按照时间顺序写成矩阵形式则为

$$X_{Tp\times q}=\begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,q} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,q} \\ \vdots & & & \\ x_{p,1} & x_{p,2} & \cdots & x_{p,q} \\ \vdots & & & \\ x_{Tp,1} & x_{Tp,2} & \cdots & x_{Tp,q} \end{bmatrix}, \tag{6}$$

即为一个 $Tp\times q$ 的矩阵，不过是按照时间先后序列按照行进行排列。

那么，根据我们本文提出的四种模式：按月竖排处理、按月横排处理、按季竖排处理、按季横排处理四种类型的数据，不同的方式得到的矩阵分别为：

按月竖排：

$$X_{mon,Tp\times q} = X_{Tp\times q}, \tag{7}$$

按月横排：

$$X_{mon,p\times Tq} = \begin{bmatrix} x_{1,1}\ x_{1,2}\ \cdots & x_{1,q}\ x_{p+1,1}\ \cdots & x_{(T-1)p+1,q} \\ x_{2,1}\ x_{2,2}\ \cdots & x_{2,q}\ x_{p+2,1}\ \cdots & x_{(T-1)p+2,q} \\ \vdots & & \vdots \\ x_{p,1}\ x_{p,2}\ \cdots & x_{p,q}\ x_{2p,1}\ \cdots & x_{Tp,q} \end{bmatrix}, \tag{8}$$

按季竖排：

$$X_{se,Tp\times q} = X_{Tp\times q}, \tag{9}$$

按季横排：

$$X_{se,p\times Tq} = X_{mon,p\times Tq}. \tag{10}$$

可见，按月竖排处理和按季竖排处理具有相同的矩阵结构形式，按月横排处理和按季横排处理也具有相同的矩阵结构形式。不同的是(见本文 3.3 部分)按月处理与按季处理时数据的分级不同。把数据(这里我们不仅采用大连地区多年的气象观测数据，同时我们测试了沈阳、青岛、北京等三个城市的多年数据)代入到不同的各自矩阵形式中，进行 SVD(奇异值分解)分解得到如下结构(其他矩阵类似，省略)：

$$X_{mon,Tp\times q} = U_{mon,Tp\times Tp}\begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_{r1} \end{bmatrix} V'_{mon,q\times q}. \tag{11}$$

在上述分解式(11)中取门限阈值(本文均取 0.2)进行奇异值的阈值化处理，然后得到不同模式下的秩以及其他的参数，最终根据式(5)计算多个地方多年气象观测资料(沈阳、青岛、北京、大连四个城市历史资料)下的平均值，得到如下表结果(表 17)：

所以，从理论分析及实验上来说按月处理(横排)对于气象数据矩阵的填充效果最佳。

6. 仿真试验

在前文中我们将大连海区历史天气数据分类、数字化，转化成矩阵变量，然后利用 SVT 矩阵填充算法对四种不同结构的数据在不同的采样率条件下依次进行了矩阵填充，当采样率在临界采样率以上时，填充效果是较为理想的，此时天气数据的相对填充误差控制在 10%以内，这符合气象数据的误差要求。为了验证程序以及算法的适用性和填充准确度，我们选取最优数据结构按月处理(横排)来验证之。

Table 17. The comparison of restored measures (Average)
表 17. 矩阵恢复测度比较(平均值)

数据结构	按月处理(竖排)	按月处理(横排)	按季处理(竖排)	按季处理(横排)
大连多年数据矩阵恢复测度 ξ 均值	1.99	1.30	2.81	2.04
北京多年数据矩阵恢复测度 ξ 均值	1.98	1.41	2.76	2.17
沈阳多年数据矩阵恢复测度 ξ 均值	1.94	1.32	2.81	2.13
青岛多年数据矩阵恢复测度 ξ 均值	2.01	1.41	2.78	2.10
平均矩阵恢复测度 ξ	1.98	***1.36***	2.79	2.11

目前我们是在已经掌握 2011 年至 2015 年大连海区所有数据的基础上，对数据进行随机采样来进行实验的，而在实际情况下，正是因为数据缺失，我们才会利用各种方法将数据补回来。因此，为了方便程序的适用性，我们进行了仿真检验试验，见图 9、图 10。

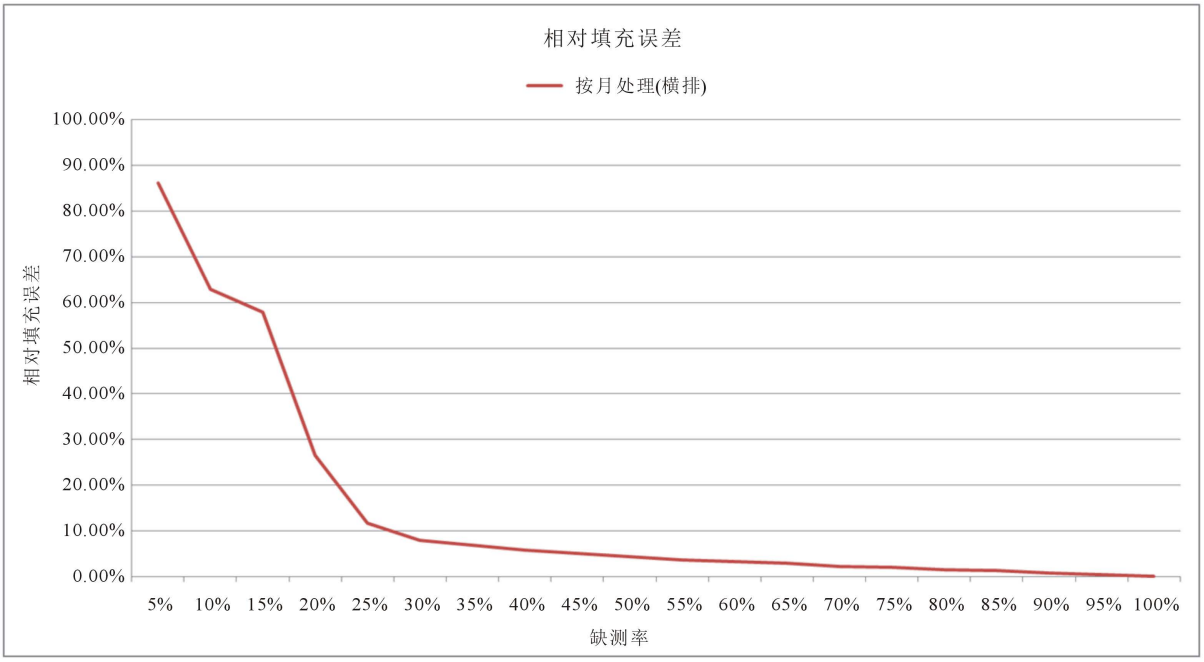

Figure 9. The error rate under different sampling rate
图 9. 不同采样率下的误差率

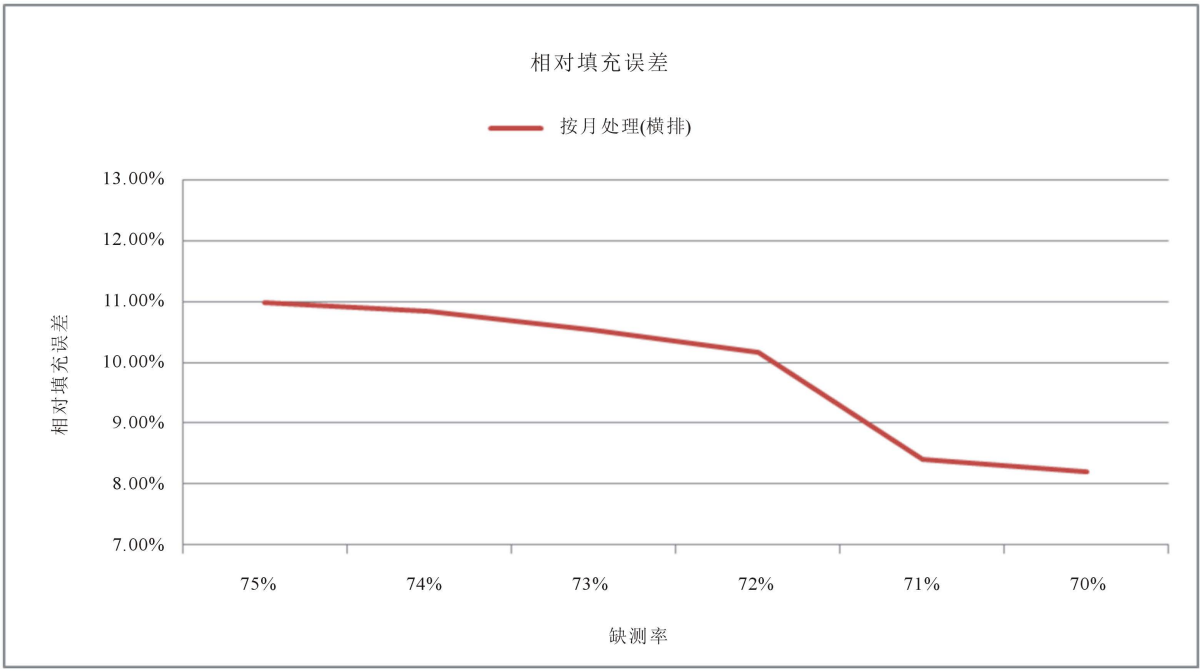

Figure 10. The error rate under different sampling rate (refined)
图 10. 不同采样率下的误差率(细化)

由图可以看出，临界缺测率约为72.10%，此时误差为9.81%。我们将缺测率72.10%对应到原始矩阵和填充矩阵。

7. 结论

本文利用大连海区2011年至2015年5年共计1825天的历史观测资料，对只有部分观测资料的情况下进行数据的恢复，并且运用历史资料进行了仿真检验试验，得到如下结论：本文模拟由于观测手段和观测背景的限制，再加上环境复杂，致使天气预报工作处于信息不全、数据缺少的状态，为缺失数据下的天气预报方法的研究提供了条件。本文所做的一部分探索是从气象数据的关联性出发，对其分类并进行数字化，实验结果表明，只要气象数据矩阵是低秩的，就可以从一小部分的关联性信息恢复得到整个的关联性信息，即一小部分关联性信息就能包含整个信息。从一系列的实验数据来看，如果有一部分相关度的气象数据，这一部分数据又是充分随机的，只要它们的数量能够达到一定的比例，即达到临界采样率，便可以通过它们恢复得到绝大部分缺失的气象数据。

利用大连海区历史资料对矩阵填充的SVT算法进行了仿真检验试验，结果表明：1)当缺测数据比例低于临界缺测率时，矩阵填充恢复的相对误差便可以控制在10%以内，这满足气象数据的使用要求，说明基于气象数据先验结构特性和矩阵优化填充技术对于气象数据补全具有很好的实用性；2)实验分析结果表明，大连地区气象观测资料月相关性明显强于季度相关性。

但是，本文在对数据处理方面仍有一定的改进空间，比如对数据的分类，文中为降低各项数据的种类，选择了“恶劣天气”、“大风”等这样的选择标准，致使某些真实情况有所弱化；另外，将类似的天气类型化为一类，比如将“小雨”和“暴雨”均改写为天气“雨”也会增大最终结果的误差。本文采用的历史数据只有大连海区等几个地区2011年至2015年5年的数据，我们都知道样本比例越高填充误差越小[15] [16] [17] [18]，所以扩充样本数量需要进一步探讨。

基金项目

国家自然科学基金(批准号：61471412，61771020，61273262)项目资助。

参考文献

[1] 朱乾根. 天气学原理和方法[M]. 南京: 南京气象出版社, 1992.

[2] 刘园园. 快速低秩矩阵与张量恢复的算法研究[D]: [博士学位论文]. 西安: 西安电子科技大学, 2013.

[3] 王萍, 蔡思佳, 刘宇. 基于随机投影技术的矩阵填充算法的改进[J]. 计算机应用, 2014, 34(6): 1587-1590.

[4] Donoho, D.L. (2006) Compressed Sensing. *IEEE Transactions on Information Theory*, **52**, 1289-1306. https://doi.org/10.1109/TIT.2006.871582

[5] 王会敏. 矩阵填充理论及其研究进展[J]. 绍兴文理学院学报, 2013, 33(7): 22-24.

[6] 黄嘉佑. 气象统计分析与预报方法[M]. 北京: 气象出版社, 2004.

[7] 王卓峥, 贾克斌. 矩阵填充与主元分析在受损图像配准中的应用[J]. 北京: 北京工业大学, 2013.

[8] Cai, J.F., Candes, E.J. and Shen, Z.W. (2010) A Singular Value Thresholding Algorithm for Matrix Completion. *SIAM Journal on Optimizatinon*, **20**, 1956-1982.

[9] Ma, S., Goldfarb, D. and Chen, L. (2008) Fixed Point and Bregman Iterative Methods for Matrix Rank Minimization. *Technical Report.*

[10] 王永曦. 矩阵填充应用于文本分类的一些探索[D]: [硕士学位论文]. 北京: 清华大学, 2012.

[11] 郭慧杰, 赵保军. 基于矩阵填充的小波图像压缩算法[J]. 系统工程与电子技术, 2012, 34(9): 1930-1933.

[12] 孟繁驰, 李书琴, 蔡骋基. 基于核范数凸优化的微阵列缺失点重建[J]. 计算机工程与设计, 2013, 34(2): 660-664.

[13] Keshavan, R.H., Montanari, A. and Sewoong, O.H. (2010) Matrix Completion from a Few Entries. *IEEE Transactions*

on Information Theory, **56**, 2980-2998. https://doi.org/10.1109/TIT.2010.2046205

[14] Benjamin, R. (2009) A Simpler Approach to Matrix Completion. *Journal of Machine Learning Research*, **12**, 3413-3430.

[15] 彭义刚, 索津莉, 戴琼海, 徐文立. 从压缩传感到低秩矩阵恢复:理论与应用[J]. 自动化学报, 2013, 39(7): 981-994.

[16] Xu, G., Wang, X., Xu, X. and Zhou, L. (2016) Entropic Uncertainty Inequalities on Sparse Representation. *IET Signal Processing*, **10**, 413-421. https://doi.org/10.1049/iet-spr.2014.0072

[17] 徐冠雷, 王孝通, 周立佳, 邵利民, 刘永禄, 徐晓刚. 广义测不准原理中的数学问题研究[J]. 应用数学进展, 2016, 5(3): 536-559.

[18] Xu, G., Wang, X., Xu, X., Zhou, L. and Liu, Y. (2017) Unified Framework for Multi-Scale Decomposition and Applications. *IET Journal of Engineering*, **2017**, 577-588.