6.2 综合误差实验

ILD 数据上的综合误差实验侧重于改变误差。同样，在其他数据集 GPS 中也观察到类似的结果，因此被省略。

6.2.1 各种错误的评估 我们考虑 Shift 和 Innovational 错误 [2, 23]，如图 8 所示的示例，其中 Spike 错误被认为是#连续错误 = 1 的特殊情况。报告相应的准确度结果在图 9 中。通常，在创新和移位错误中观察到类似的结果。也就是说，虽然我们提出的 IMR 处理尖峰错误（即图 9 中的 # 个连续错误 = 1）不如 SCREEN [22] 好，但 IMR 在处理大量问题时总是显示出明显更好的结果（更低的 RMS 测量）在 Shift 和 Innovational 错误模式上的连续错误。结果再次证明我们的建议在修复连续错误方面效果很好。如图 8(a) 所示，当时间序列移回非异常数据时，不仅提出的 IMR，而且具有小阈值 τ 的 ARX 都会遭受过校正。附在每种方法上的数字，例如 IMR-0.05，表示 IMR 的 τ = 0.05。如图所示，在 IMR 和 ARX 中都有一个权衡：较小的阈值 τ 在处理连续错误时显示出更好的结果，但在移位结束时会导致过度校正。然而，如图 5 和 11 所示，较小的阈值 τ 通常具有更好的整体精度。还值得注意的是，当时间序列转移回非异常数据时，现有方法 EWMA 平滑和 SCREEN 也无法很好地处理修复，如图 1 所示的示例。

6.2.2 在线计算评估

只要在线标签可用（在第 1.3 节中讨论），建议的 IMR 就适用。值得注意的是，通过将所有历史数据解释为一个标记段第 4.2.2 节中无需迭代的直接修复计算可以应用于高效的在线计算。图 12 显示了在线修复的结果。由于错误是在数据集中随机引入的，为了获得可靠的结果，我们对每个引入随机错误的测试重复 10 次并报告平均值。 IMR 的吞吐量是稳定的，与其他的可比。 IMR 再次显示最佳修复。与 ARX 相比，IMR 的改进不如其他实验显着。这并不奇怪，参考 ARX 的类似公式 3 和 IMR(1) 的公式 10。 IMR 的优点是 IMR(1) 在 4.2.2 节中有一个标记段不需要阈值参数，而 ARX 需要设置阈值 τ。

6.2.3 温度实验

为了评估循环时间序列，我们使用了另一个以年为单位的温度数据集（http://data.cma.cn，循环模式如图 13 所示）。与 ILD 类似，我们将合成误差注入到温度数据中。图 14 显示了不同错误类型的结果。同样，结果通常与图 9 中的 ILD 相似。也就是说，所提出的 IMR 在处理大量连续错误时显示出明显更好的结果。 SCREEN 在 Spike 错误（即，#连续错误 = 1）下在该数据集中没有显示更好的结果，因为干净的数据还包含大量的 Spike，如图 13 所示，并且无法与错误区分开来。

1. 相关工作
2. 还研究了在多次迭代中执行修复的想法[24]，过去的修复可能有助于在未来推荐更准确的修复。然而，连续数据清理方法 [24] 并不直接适用于我们的问题，因为它采用了 FD 约束，而这在时间序列数据中是不可用的。 IMR SCREEN IMR-MP EWMA IMR-IC ARX AR 0.1 1 10 1 5 10 15 20 25 30 RMS 错误 # 连续错误 (a) 0.1 1 10 1 5 10 15 20 25 30 RMS 错误 # 连续错误 (b) 图 14：不同数量的连续错误，在 (a)Shift 和 (b)Innovational 错误类型下，在 τ = 0.2、p = 3 和数据大小 1.5k 的温度范围内

7.1 时间数据上的异常检测

AR 和 ARX 确实已广泛用于异常在经济和社会调查等各个领域进行检测 [3, 6]。我们在本研究中考虑了 ARX [19]，因为这种方法可以利用标记的事实，并且可以适应数据修复中的最小变化原则（如第 3 节所示）。 Hellerstein [13] 调查了大型数据库定量属​​性中清理错误的方法，其中时间序列数据也作为一种特殊情况进行了讨论。例如，Tsay [23] 提出了用于检测和处理单变量时间序列中的异常值和结构变化的统一方法。迭代过程包括规格估计检测去除周期以逐一处理最显着的干扰。 ARIMA [18, 3] 是时间序列的一般参数族，由自回归过程和移动平均过程组成。它还可以包含范围广泛的非平稳序列。

7.2 基于平滑的清理

平滑技术通常用于消除噪声数据。例如，简单移动平均线 (SMA) [4] 通过计算最后 k 个点的未加权平均值来平滑时间序列数据。指数加权移动平均线 (EWMA) [10] 不是平均加权，而是随着时间的推移分配指数下降的权重。如[22]所示，也如示例2中的图2所示以及实验中的图3（EWMA）所示，平滑方法可能会严重改变原始正确数据，因此修复精度较低。相比之下，我们基于最小变化的方法，仅应用高置信度修复，可以以更好的修复精度保留大部分原始值。

7.3 基于约束的清洗 基于约束的修复被广泛用于清洗脏数据，使得修复后的数据满足一些给定的约束，并且最小化修复修改[1, 8]。为了清理顺序数据，现有研究 [22] 采用了一类速度约束，声明值变化的速度应该是有界的。因此，修复是为了满足这种速度限制而修改序列。这种基于约束的修复在两个方面存在不足：（1）它不能处理一系列连续的错误，（2）没有利用标记的事实。

8.结论在本文中，我们研究了修复脏时间序列数据的问题，给定了一些数据点的标签真实性。 (1) 虽然现有的异常检测技术可以适用于修复，但我们认为基于显着偏差（观察和预测之间）的异常检测与数据修复中的最小变化原则不一致。我们对真实数据集的实验说明了在异常修复中应用异常检测的这种不一致性。 (2) 我们由此提出了一种迭代最小修复 (IMR) 算法。通过在错误预测的每次迭代中创造性地执行一次最小修复，该算法将异常检测中捕获时间性质的美与数据修复中的最小变化结合在一起。再次，实验证明了我们提议的优越性。值得注意的是，与异常检测方法 AR 和 ARX 相比，我们提出的 IMR 对阶 p 的设置不敏感，即，小的 p 足以以低时间成本实现高修复精度。 (3)明确分析了IMR的收敛性。特别是，我们表明在某些情况下可以直接计算收敛修复结果而无需迭代计算，从而可以对流数据进行有效的在线修复。值得注意的是，与现有的 ARX 不同，在线计算中的 IMR 不需要指定阈值。 (4) 最后，我们设计了高效的剪枝和增量计算，将参数估计的复杂度从线性时间降低到恒定时间。实验表明通过剪枝和增量计算可以显着提高时间性能。