1876

HENDRIX
COLLEGE

Data Challenge 1

Mathematical Models

Author

Paul Beggs
BeggsPA@Hendrix.edu

Instructor

Dr. Christopher Camfield, Ph.D.

Due

February 10, 2026

1 Introduction

In the present problem, our goal was to develop a strategy for calculating the average temperature of a 2D surface plotted on an xy-plane with uneven sampling. We accomplished this goal by exploring the efficacy of different strategies and how viable they would be at solving the problem. These strategies are: the **Naïve Approach** and **Grid Partitioning**. The naïve strategy assumes a uniform distribution of all data points and averages all points regardless of concentration. Conversely, by partitioning, evaluating, and allocating, we could subdivide the plane into different granularities (e.g., four partitions versus eight). With this strategy, we were able to quantify each region's representation in the overall average and justify the resulting average temperature.

2 Strategies

2.1 Naïve

The naïve strategy consists of assuming all points on the surface are spread equally among themselves, and then averaging by taking the sum of the temperature values and dividing by the number of points (counts). This approach is inadequate for the present problem because we have high concentrations of data toward the leftmost part of the plane. Put concretely, there are 21 data points within the box $x \in [15, 20]$, $y \in [0, 10]$, which constitutes 42% of the available data. Thus, we need a more sophisticated approach that scales the over-sampled regions to match sparsely-sampled ones.

2.2 Partition, Evaluate, & Allocate

In this approach, we partitioned all points into exclusive groups, averaged each group's temperature with their respective counts, and then calculated a representation factor to assess the sampling bias. Considering how many partitions to include was a challenge. We wanted to find the balance between too many partitions and too few. At first, we tried four partitions, and this gave an average temperature of 156.1 (compared to the naïve temperature of 151.0). Then, when we tried eight partitions, we calculated a temperature of 155.8.

However, knowing the temperatures was not enough; in order to rank them, we needed to calculate another statistic: the sampling bias factor for each region. This factor was calculated by dividing the effective weight of the data in the average (e.g., the 42% mentioned earlier) by the partition weight (e.g., 12.5% for eight partitions). This gives a ratio that essentially just tells us how "good" the partition data is to what a fair average would be. If the ratio is close to one, then we know that the sampled data is weighted appropriately. If the data is less than one, then the data is under-represented by the naïve average; conversely, if it is greater than one, then the data is over-represented. Consequently, it is advantageous to subdivide until we reveal the true extremes of under-represented regions ($.50 \leq$) to counteract the greatly over-represented ones (≥ 1.50), rather than letting large partitions mask disparities by averaging them out. This is exactly what we found with 8 partitions: three regions of under-representation compared to 2 of over-representation.

3 Conclusion

Through this exercise, we uncovered that it is possible to still get a representative average of a plane, even when the data is messy. We accomplished this goal by partitioning the data into representative chunks, which were quantified with a sampling bias factor. This told us that our solution appropriately weighted each region, so that none were more important than the others.