Inpainting of Radio Interferometric Measurements: Decorrelation and Baseline-dependent Weighted Nearest Neighbours Imputation

1. **Broad area of research:** Science

2. **Academic level of research project:** PhD

3. **Abstract:** Radio Frequency Interference (RFI) remains a critical challenge for modern radio telescopes like MeerKAT, causing significant data loss through flagging that can bias scientific results and mask transient events. Traditional approaches simply exclude flagged data, leading to gaps in frequency and time coverage that degrade signal-to-noise ratios and introduce systematic errors in spectral analysis and transient detection. This project develops a novel baseline-dependent K-Nearest Neighbors (KNN) algorithm to predict and recover flagged visibility data in MeerKAT observations. Unlike standard KNN methods that use Euclidean distance metrics, our approach incorporates baseline-specific distance calculations that account for how different projected baseline lengths sample the uv-plane during aperture synthesis.

4. **Primary supervisor:** Prof Oleg Smirnov, o.smirnov@ru.ac.za, Rhodes University & SARAO & INAF-IRA

5. **Co-supervisor:** Prof Marcellin Atemkeng, m.atemkeng@ru.ac.za, Rhodes University

Details of Research Project

1. Scientific merit

In recent years, we have witnessed the construction of several new radio telescopes designed with specific scientific capabilities. Notable projects include the Square Kilometre Array (SKA) precursors at high and low radio frequencies such as the MeerKAT telescope, the Low-Frequency Array (LoFAR), and the Murchison Widefield Array (MWA). One of the challenges that significantly affects the performance of these instruments is the corruption of their measurements, primarily caused by Radio Frequency Interference (RFI). RFI poses a significant challenge in radio astronomy, primarily due to its diverse sources, including terrestrial transmitters like radar signals and telecommunication systems. This interference typically leads to narrowband flagging during the initial stages of data preprocessing. Additionally, there are wideband RFI sources, such as satellite systems, that require the identification and flagging of substantial portions of the raw data. Removing RFI during data analysis is crucial for upholding the precision of astronomical observations. However, when the RFI is removed, it creates gaps in the data, which can potentially negatively affect science and the functionality of the software used to analyze the data.

Data flagging can also occur due to a range of issues, including instrument errors, cosmic rays, or other sources of artifacts. Dealing with flagged data in radio interferometric data presents numerous challenges that require careful consideration, expertise in data analysis techniques, and ongoing efforts to improve data quality and data handling processes. One must carefully consider how to treat flagged data points, whether to exclude them entirely or apply corrections that maintain data integrity while minimizing biases. There is a risk of losing both frequency channels and timeslot data, which are typically excluded during the analysis and potentially lead to biased results and a decrease in signal-to-noise (SNR) of the observation. For example, as a result of RFI removal, there is a potential risk of short transient events going undetected due to a flagged data stream that contains transients before the data is processed for transient analysis. Using a LOFAR survey in their investigations, Cendes et al. (2018) highlighted that transient events with periodicities close to the time of a single timeslot or few are entirely flagged.

On the other hand, if the flagged values are replaced with predictions via some sort of interpolation or extrapolation, we risk introducing uncertainty and assumptions that can affect the validity of the results. In Chakraborty et al. (2022), a comparative analysis was conducted to address the issue of missing spectral information caused by RFI in cosmological HI 21 cm observations. The results show that in the presence of periodic RFI flagging as well as periodic plus broadband RFI flagging, the CLEAN method failed to effectively mitigate foreground leakage, which introduced an additional bias to the estimated power spectrum. Flagged data can highlight issues with instruments or data processing pipelines. Addressing these issues may require instrument modifications, software updates, or improvements in data reduction techniques.

This work will investigate using KNN to predict the flagged samples. The KNN algorithm is an intuitive supervised machine learning algorithm commonly used for classification and regression tasks. It operates on the principle of proximity, where data points that are close to each other in feature space are assumed to belong to the same class or exhibit similar properties. The prediction starts by calculating the distance between the data point to be predicted and all data points in the training dataset. Commonly used distance metrics include the Euclidean distance, Manhattan distance, or Minkowski distance. However, the choice of distance metric can be modified to the specific problem. In this work, our first contribution is to modify the KNN distance metric to be baseline-dependent which captures the variation and orientation of each projected baseline during the aperture synthesis. The algorithm selects the k-nearest neighbors based on the shortest distances and computes the average (or weighted average) of these k-nearest neighbor data points. The resulting average value is then used as the prediction of the missing value. Selecting the appropriate value of k is crucial. A smaller k value can result in a more sensitive model, which is prone to noise. A larger k can provide a smoother prediction but may lead to decorrelation.

Some of the contributions of this work will be as follows

- Baseline-dependent window functions will be used with a large value of k to mitigate decorrelation while making the prediction more sensitive to the target value.
- The distance metric will be calculated to consider the variation in how visibility data is sampled across baselines in the uv-space. Longer east-west projected baselines cover larger distances in the uv space compared to shorter east-west projected baselines.
- k_{pq} is baseline-dependent (does the subscript pq), while using baseline-dependent distances and baseline-dependent window functions to standardize the distance scale.

Link to SARAO research priority areas for 2026: This is a science project using existing MeerKAT data, thus maximizing the science yield of the existing archive.

2. Feasibility: MeerKAT experiences RFI contamination despite radio quiet zone protections. Common sources include satellite communications, aviation radar, and mobile telecommunications, resulting in both narrowband and wideband flagging. This challenging RFI environment makes MeerKAT an ideal testbed. MeerKAT has accumulated archival datasets since commissioning, including continuum surveys, HI surveys, and pulsar timing observations. These diverse datasets span multiple observing modes (spectral line, continuum, time-domain), allowing comprehensive testing of the baseline-dependent KNN across different scientific use cases. The availability of repeated observations of calibrator sources provides ground truth for validation; comparing RFI-recovered data against clean observations of the same source.

3. Data availability and equipment access. This project is entirely based on archival public MeerKAT data, available via the SARAO interface. The student will have full access to the RATT compute cluster, which is well-equipped to handle the project's data processing and storage requirements. Additional compute resources are available on the ILIFU cluster, if required.

4. Project timeline
Months 1 – 6: Literature Review, Algorithm Development, and Data Preparation.

Literature review of RFI mitigation techniques in radio interferometry, including traditional flagging methods, interpolation approaches, and machine learning applications. Study of KNN algorithms, baseline-dependent weighting schemes, and distance metrics in interferometric data analysis. Development of baseline-dependent distance metric formulation, incorporating uv-space sampling variations and baseline orientation. Implementation of baseline-dependent window functions to address decorrelation.

Months 6 – 12: Algorithm Validation and Parameter Optimization

Processing of initial MeerKAT calibrated datasets with standard flagging pipelines to establish baseline RFI statistics and flagging patterns. Implementation of synthetic RFI injection across different RFI types. Application of baseline-dependent KNN to synthetic RFI scenarios, computing recovery accuracy metrics. Comparative analysis against standard KNN with

Euclidean distance and simple interpolation methods. Cross-validation using repeated MeerKAT observations of the same calibrator sources.

Months 12 – 24: Application to Science Cases and Comprehensive Testing

Systematic application of optimized baseline-dependent KNN to diverse MeerKAT science datasets. Quantitative comparison of science outcomes: imaging dynamic range, spectral fidelity, transient detection completeness, and computational efficiency. Analysis of failure modes and identifying conditions where KNN recovery degrades performance.

Preparation of first paper: "Baseline-Dependent K-Nearest Neighbors for RFI Mitigation in Radio Interferometry: Application to MeerKAT".

Months 24 – 36: Publication and Thesis Completion

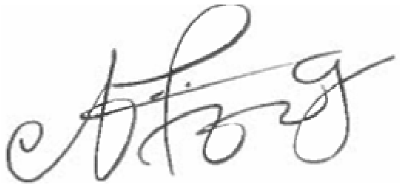
Submission of primary methodology paper to A&A or MNRAS. Development of open-source software package implementing baseline-dependent KNN, compatible with standard radio astronomy tools. Documentation and tutorial notebooks for adoption. Presentation of results at conferences (e.g., SARAO Postgraduate Conference, URSI General Assembly, SKA Science Meeting).

5. **Qualifications, academic abilities, skills and/or experience that a student should have in order to successfully deliver on the objectives of the research proposed:** Good knowledge of Python programming. Familiarity with radio interferometry and radio data reduction would be advantageous, but not required.

Supervisor

Oleg Smirnov 26 February 2026

Co-Supervisor



Marcellin Atemkeng 26 February 2026