Mining the GPIES Database

[SPIE: 10703-17]

Dmitry Savransky
Jacob Shapiro, Vanessa Bailey, Robert De Rosa, Jason Wang,
Jean-Baptiste Ruffio, Eric Nielsen, Melisa Tallis, Marshall Perrin
and the GPIES Team

June 10, 2018
The Gemini Planet Imager Exoplanet Survey

484 targets observed to date

Figures courtesy of R. De Rosa
GPIES Campaign Data System

About 3 years into the survey, we have accumulated Δ∼ 61 million contrast curve data points. In addition to the science datacubes, ∼ 5600 stars themselves. To handle all of the data while also making it available to the entire collaboration, we use a combination of collaborative tools such as Slack and our internal wiki are integrated into our automated data infrastructure (Sec. 2.1.1). We also describe how observing tools (Sec. 2.4) demonstrate the sensitivity achieved in a single frame of data. These quicklook reductions allow observers to assess data quality so that they are able to recognize and flag poor-quality data. In about 3% of the process, the observer can log that particular file as bad through the GPItv interface in the GPI DRP. The summit quicklook reductions provide observers the basic tools to assess data quality so that bad data can be identified and excluded from future analysis.

After new IFS data are taken on the summit, an instance of the GPI DRP running on the summit uses the GPI DRP autoreducer module to automatically perform quicklook reductions of the reprocessed data, we have generated ∼ 68 million contrast curve data points. During an observing sequence, the observer can log that particular file as bad through the GPItv interface in the GPI DRP. The summit quicklook reductions provide observers the basic tools to assess data quality so that bad data can be identified and excluded from future analysis.

To move the data off the summit, automated scripts download after some data processing has happened such as whether the data product is a quicklook or science-grade reduction, the observer can log that particular file as bad through the GPItv interface in the GPI DRP. The summit quicklook reductions provide observers the basic tools to assess data quality so that bad data can be identified and excluded from future analysis.

We also describe how observing tools (Sec. 2.4) demonstrate the sensitivity achieved at some fiducial separations, flux calibration conversions, as well as a unique ID for the reduced file. To link the raw data to their original raw data products, a third database table is a two-column table where each row associates one raw file ID with one reduced file ID. Multiple reduced data tables contain one row for each reduced file, produced either by or. The reduced data table contains information on the target wavelength band, time of observation, along with a column for each of the fields in the file headers (e.g., observing mode, observing can continue in the unlikely case the observatory network is ineffective. During an observing sequence, the observer can log that particular file as bad through the GPItv interface in the GPI DRP. The summit quicklook reductions provide observers the basic tools to assess data quality so that bad data can be identified and excluded from future analysis.

From: [Wang et al., 2018]
GPIES Database Schema

- Observation Notes
- Raw Data Products
  - Targets
    - GPIES Objects of Interest
    - Target Property Tables
- Reduced Data Products
  - Raw2Reduced
- Contrasts
  - Contrasts2RID
- AO Raw Files
  - AO Raw2Reduced
  - AO Reduced
  - AO Raw
GPIES Database Contents

- 14 GB (metadata and ancillary products only)
- 136,151 IFS Raw Data Files (30,142 GPIES)
- 263,973 IFS Reduced Data Products
- 86,092 AO Raw Telemetry Files
- 86,325,374 Contrast Values

What can we do with all this data?
GPIES Database Contents

- 14 GB (metadata and ancillary products only)
- 136,151 IFS Raw Data Files (30,142 GPIES)
- 263,973 IFS Reduced Data Products
- 86,092 AO Raw Telemetry Files
- 86,325,374 Contrast Values

What can we do with all this data?
Previous Work

- Performance characterization of GPI’s AO with IFS data
  [Poyneer et al., 2016, Bailey et al., 2016]
- GPI performance variation characterization with operating conditions
  [Tallis et al., 2018]
- See also: Tallis et al., this conference [10703-267]
Previous Work

- Performance characterization of GPI’s AO with IFS data
  [Poyneer et al., 2016, Bailey et al., 2016]
- GPI performance variation characterization with operating conditions
  [Tallis et al., 2018]
- See also: Tallis et al., this conference [10703-267]

That’s not what this talk is about

Here, we are only looking for purely data-driven results, with no specific physical modeling of underlying processes
Finding Correlations

For two random variables $\bar{x}, \bar{y}$:

- **Pearson product-moment**:
  \[
  r_{\bar{x}, \bar{y}} = \frac{E[(\bar{x} - \mu(\bar{x}))(\bar{y} - \mu(\bar{y}))]}{\sigma(\bar{x})\sigma(\bar{y})}
  \]

- **Spearman rank correlation**:
  \[
  \rho_{\bar{x}, \bar{y}} = r_{\text{rank } \bar{x}, \text{rank } \bar{y}}
  \]

- **Kendall rank correlation**:
  \[
  \tau = \frac{2}{n(n-1)} \left( \sum_{i \neq j} [((x_i > x_j) \& (y_i > y_j)) | ((x_i < x_j) \& (y_i < y_j))] 
  - \sum_{i \neq j} [((x_i < x_j) \& (y_i > y_j)) | ((x_i > x_j) \& (y_i < y_j))] \right)
  \]
Contrast Correlations

0.40 arcsec T-Type Contrast

0.40 arcsec L-Type Contrast
Contrast Correlations

0.80 arcsec T-Type Contrast

0.80 arcsec L-Type Contrast
The Data is Noisy
A point $y_i$ may belong to the “true” data or be considered an outlier drawn from a normal distribution $\sim (\mu_o, \sigma_o)$, governed by binary flag $o_i$:

$$p(y_i|x_i, \sigma_i, \theta, o_i, \mu_o, \sigma_o) = \frac{1}{\sqrt{2\pi (\sigma_i^2 + o_i\sigma_o^2)}} \exp \left( - \frac{[y_i - (1 - o_i)f_\theta(x_i) - o_i\mu_o]^2}{2(\sigma_i^2 + o_i\sigma_o^2)} \right)$$

The marginalized likelihood is then:

$$p(\{y_i\}_{i=1}^n | \{x_i\}_{i=1}^n, \{\sigma_i\}_{i=1}^n, \theta, \mu_o, \sigma_o) = \prod_{i=1}^n \left[ Op(y_i|x_i, \sigma_i, \theta, o_i = 0) + (1 - O)p(y_i|x_i, \sigma_i, \theta, o_i = 1) \right]$$

for

$$p(o_i) = \begin{cases} O & o_i = 0 \\ 1 - O & o_i = 1 \end{cases}$$

See: [Hogg et al., 2010, Hogg and Foreman-Mackey, 2017]
Linear Modelling (I-Magnitude)

0.25 as

0.40 as

0.80 as
Linear Modelling (Ambient Temperature)

0.25 as

0.40 as

0.80 as
DNN Regression

From: [G. Lion, 2016]

This work done entirely in TensorFlow r1.8.
Choice of Network

![Graph showing the relationship between Neurons and Testing Mean Square Error. The x-axis represents the number of Neurons (4 to 10), and the y-axis represents the Testing Mean Square Error (0.175 to 0.350). The graph shows an upward trend as the number of Neurons increases.]
Single Layer, 8 Neuron, 9 Input Regression Network

RMSE: 0.40
Two Layers, 16 Neuron, 6 Input Regression Network

RMSE: 0.18
What Can We Say After the First Observation?
Three Layers, 60 Neuron, 22 Input Regression Network

\[
\begin{array}{c|c|c|c|c|c}
 & 5.0 & 4.5 & 4.0 & 3.5 & 3.0 \\
\hline
\text{True Value} & 5.0 & 4.5 & 4.0 & 3.5 & 3.0 \\
\text{Predicted Value} & 5.0 & 4.5 & 4.0 & 3.5 & 3.0 \\
\end{array}
\]

RMSE: 0.11
Conclusions

- Jointly exploiting operational and science data metrics can lead to new discoveries, but is difficult if you don’t have the proper infrastructure in place.
- Polynomial models are likely insufficient to accurately describe performance variations given the large numbers of endogenous and exogenous factors in play.
- Machine Learning is great, but it’s hard to tell if you really have the right answer.
References I


