PNNL-20332

U.S. DEPARTMENT OF ENERGY

Prepared for the U.S. Department of Energy
under Contract DE-AC05-76RL01830

Application of Bayesian Dig-Stop Algorithm to Camp San-Luis Obispo Classification Demonstration Data

Stephen Walsh
John Hathaway
Kevin Anderson
Brent Pulsipher

April 2011

Pacific Northwest
NATIONAL LABORATORY

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor Battelle Memorial Institute, nor any of their employees, makes **any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights**. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or Battelle Memorial Institute. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

PACIFIC NORTHWEST NATIONAL LABORATORY
operated by
BATTELLE
for the
UNITED STATES DEPARTMENT OF ENERGY
under Contract DE-AC05-76RL01830

Printed in the United States of America

Available to DOE and DOE contractors from the
Office of Scientific and Technical Information,
P.O. Box 62, Oak Ridge, TN 37831-0062;
ph: (865) 576-8401
fax: (865) 576 5728
email: reports@adonis.osti.gov

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

The public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing the burden, to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.
PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
04/01/2011	Technical	April 2011

4. TITLE AND SUBTITLE

Application of Bayesian Dig-Stop Algorithm to Camp San-Luis Obispo Classification Demonstration Data

5a. CONTRACT NUMBER

PNNL-20332

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Stephen Walsh
John Hathaway
Kevin Anderson
Brent Pulsipher

5d. PROJECT NUMBER

MR-200837

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Pacific Northwest National Laboratory
P.O. Box 999
Richland, WA 99352

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Cleared for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

With millions of acres of DoD ranges being considered for other uses, unexploded ordnance continues to pose a significant challenge. For heavily used target areas, the number of anomalies identified can be very large. One of the largest costs during the remediation process is the cost associated with digging every anomaly. To address this, ESTCP embarked on a program to enhance and demonstrate technologies for discriminating between MEC and non-MEC buried objects. For this demonstration at a site located in San Luis Obispo, CA, these technologies and classification analysis routines generated a metric which reflects the degree of belief that the anomalies are Targets Of Interest (TOI). The objective is to use these metrics in an informed way to allow early termination of anomaly digging, thereby reducing the overall cost of remediation while minimizing the likelihood of undug MEC to acceptable levels.

PNNL has developed a viable statistically-defensible algorithm, Bayesian Dig Stop (BDS), that could be used to support early dig stopping recommendations. It allows one to state with X% confidence that there is no more than a Y% chance that MEC remains on

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Dr. Herb Nelson
u	u	u	SAR	44	19b. TELEPHONE NUMBER *(Include area code)* 703.696.2127

Executive Summary

With millions of acres of DoD ranges being considered for other uses, unexploded ordnance continues to pose a significant challenge. The Environmental Security Technology Certification Program (ESTCP) has demonstrated wide area assessment methods and tools that can be used to identify and delineate target areas where the likelihood of Munitions and Explosives of Concern (MEC) is greatest. These delineated areas are then characterized through 100% geophysical surveys and any geophysical anomalies are identified and dug up. For heavily used target areas, the number of anomalies identified can be very large. One of the largest costs during the remediation process is the cost associated with digging every anomaly.

To reduce the cost of remediation, ESTCP embarked on a program to enhance and demonstrate technologies for discriminating between MEC and non-MEC buried objects. For this demonstration at a site located in San Luis Obispo, CA, these discrimination technologies and classification analysis routines generated a metric which reflects the degree of belief that the anomalies are Targets Of Interest (TOI). TOI's are defined as MEC anomalies or anomalies which exhibit MEC properties. These metrics are then used to determine the anomaly digging order. The metrics are unable to exactly classify each anomaly as TOI/Not-TOI due to measurement error (including machine noise and environmental nuisance factors). The objective is to use these metrics in an informed way to allow early termination of anomaly digging, thereby reducing the overall cost of remediation while minimizing the likelihood of undug MEC to acceptable levels. Therefore, analysts are required to identify a "conservative" threshold at which, after digging anomalies with large metrics, they believe the remaining anomalies are Not-TOI.

PNNL has developed a viable statistically-defensible algorithm, Bayesian Dig Stop (BDS), that can be used to support early dig stopping recommendations. It allows one to state with X% confidence that there is no more than a Y% chance that MEC remains on the site. The methods can be used after an initial set of digs or sequentially as digging proceeds. It also provides a universal metric to compare the performance of various anomaly classification algorithms. The main objectives of the Bayesian Dig-Stop methodology are as follows.

1. Provide justification for decision to stop digging anomalies during the remediation process
2. Quantify the degree of confidence that few or no Targets of Interest (TOI) remain undug
3. Account for past discriminating performance of the sensor system using training or historical data
4. Allow updates to the stop digging decision rule and confidence statements as digging proceeds

In this report, we document the application of BDS to 25 of the 54 dig-lists produced in the Camp San Luis Obispo (SLO) demonstration. We furthered the analysis capabilities by constructing preliminary diagnostics which indicate if the assumptions of the BDS algorithm are met by the data. The results were mixed: some dig lists yielded reasonable dig stop criteria (95% confidence that the probability of no remaining TOI is either .95 or .99) while saving a significant number of anomalies digs and recovering all or almost all TOI, but this was observed with varying degrees of meeting/violating the assumptions of BDS. For other dig lists larger confidence levels were achieved but only at the expense of recovering all anomalies (which was the correct decision). The results are summarized in Table 1.

	All columns use the 95% lower confidence bound					
	Number of anomalies undug		Proportion of anomalies undug		Number of TOI remaining	
Sensor/Demonstrator	95% probability that no TOI remain	99% probability that no TOI remain	95% probability that no TOI remain	99% probability that no TOI remain	95% probability that no TOI remain	99% probability that no TOI remain
SIG_MetalMapper_PNBC	928	647	0.65	0.46	0	0
SAIC/UXAnalyze_EM61 Array	292	204	0.22	16	1	0
SAIC/UXAnalyze_EM61 Cart No slope correction	604	592	0.47	0.46	1	1
SAIC/UXAnalyze_EM61 Cart Slope correction	605	582	0.47	0.45	1	1
SAIC/UXAnalyze_MSEMS	455	445	0.35	0.34	1	1
SIG_MetalMapper_RVM	1163	1049	0.82	0.74	7	2
SAIC_TEMTADS_2 Criteria	851	823	0.68	0.64	3	3
SAIC_TEMTADS_3 Criteria	650	547	0.51	0.42	3	3
SAIC_MetalMapper_2Criteria	1070	1052	0.76	0.74	4	3
SAIC_MetalMapper_3Criteria	1060	954	0.75	0.68	3	2
Sky_MetalMapper	1054	993	0.75	0.71	2	2
SIG/UBC_MAG_PNBC	350	205	0.30	0.18	7	2
SIG_EM61 Array_RVM	< 271	NA	< 0.24	NA	3	NA
SIG_MAG_PNBC	NA	NA	NA	0	0	0
SIG/UBC_EM61 Cart_PNBC	NA	NA	NA	0	0	0
SIG/UBC_MAG_RVM	NA	NA	NA	0	0	0
SIG/UBC_MSEMS_PNBC	NA	NA	NA	0	0	0
SIG/UBC_TEMTADS_PNBC	< 131	NA	< 0.10	NA	1	NA
SIG/UBC_TEMTADS_RVM	< 426	NA	< 0.33	NA	2	NA
SIG_EM61 Array_PNBC	42	NA	0.04	NA	0	NA
SIG_MAG_RVM	35	31	0.027	0.025	1	1/1
SIG_MSEMS_PNBC	< 204	NA	< 0.16	NA	0	0/0

Table 1: BDS summary of results as applied to the 25 SLO dig lists. Green highlighting indicates dig lists where BDS produced large 95% LB on Pr(No TOI Remaining | Data), a significant number of anomaly digs were saved, and where only 0 or 1 TOI were left unrecovered. Yellow highlights indicate dig lists where a large 95% LB was produced by BDS, where a significant number of anomaly digs were saved, but where 2 or more TOI were left unrecovered. Red highlighting indicates dig lists where BDS was unable to produce high confidence that all TOI were recovered, thus all anomalies had to be dug. White highlighting indicates dig lists where BDS measured high confidence that all TOI could be recovered at some dig threshold, but where the threshold was so low that almost all anomalies needed to be recovered.

The MetalMapper system using the SIG PNBC classification algorithm seemed to perform well and the BDS algorithm suggested that one could stop digging after digging only 35% of

the anomalies to achieve 95% confidence that the probability of no remaining TOI was at least 0.95. To achieve 95% confidence that the probability of no remaining TOI is at least 0.99, then digging could be stopped after 54% of the anomalies were dug. The performance of other sensor/algorithm combinations are summarized in Table 1.

We believe that the BDS methodology fills an important gap in the entire process of anomaly classification and elimination of unnecessary digs. The unique contributions outlined above can provide an additional level of rigor and confidence in this process. However, issues outlined in the Conclusions and Recommendations section relative to its implementation must be resolved before it will be accepted to accurately characterize confidence. Recommendations for resolving these issues include:

- Evaluation of model fitting diagnostics for the BDS methodology
- Explore the BDS algorithm's ability to take advantage of all available data and re-rank the anomalies after each major dig sequence
- Extend the BDS methodology to work with the multivariate feature space data rather than with the subjective, information-poor scoring metric.
- Adapt methodology to incorporate performance data on previous sites, given the possibility of no or limited training data on actual site.

Table of Contents

1 Introduction 1
2 Bayesian Digstop Methodology 2
2.1 BDS Objectives and Use 3
2.2 BDS Contributions to the State of the Art 4
3 BDS Application to SLO Data 5
3.1 Results Summary 11
4 Conclusions and Recommendations 13

Table of Figures

Figure 1: Flow chart depicting the major components of the BDS algorithm as applied to site remediation 3
Figure 2: SAIC Metal Mapper (3 criteria) BDS fit and diagnostics. 8
Figure 3: SIG-UBC TEMTADS (RVM) BDS fit and diagnostics. 10

Figure A 1: SIG EM61 Array (PNBC) BDS fit and diagnostics. 16
Figure A 2: SIG EM61 Array (RVM) BDS fit and diagnostics. 17
Figure A 3: SIG MAG (PNBC) BDS fit and diagnostics. 18
Figure A 4: SIG MAG (RVM) BDS fit and diagnostics. 19
Figure A 5: SIG MSEMS (PNBC) BDS fit and diagnostics. 20
Figure A 6: SIG TEMTADS (PNBC) BDS fit and diagnostics. 21
Figure A 7: SIG-UBC EM61 Array (PNBC) BDS fit and diagnostics. 22
Figure A 8: SIG-UBC EM61 Array (RVM) BDS fit and diagnostics. 23
Figure A 9: SIG-UBC EM61 Cart (PNBC) BDS fit and diagnostics. 24
Figure A 10: SIG-UBC MAG (PNBC) BDS fit and diagnostics. 25
Figure A 11: SIG-UBC MAG (RVM) BDS fit and diagnostics. 26
Figure A 12: SIG-UBC MSEMS (PNBC) BDS fit and diagnostics. 27
Figure A 13: SIG-UBC TEMTADS (PNBC) BDS fit and diagnostics. 28
Figure A 14: SAIC UXAnalyze EM61 Array BDS fit and diagnostics. 29
Figure A 15: SAIC EM61 Cart (no slope correction) BDS fit and diagnostics. 30
Figure A 16: SAIC EM61 Cart (slope corrected) BDS fit and diagnostics. 31
Figure A 17: SAIC UXANALYZE MSEMS BDS fit and diagnostics. 32
Figure A 18: SAIC EM61 TEMTADS (2 criteria) BDS fit and diagnostics. 33
Figure A 19: SAIC EM61 TEMTADS (3 criteria) BDS fit and diagnostics. 34
Figure A 20: SAIC Metal Mapper (2 criteria) BDS fit and diagnostics. 35
Figure A 21: SKY Metal Mapper BDS fit and diagnostics. 36
Figure A 22: SIG Metal Mapper (PNBC) BDS fit and diagnostics. 37
Figure A 23: SIG Metal Mapper (RVM) BDS fit and diagnostics. 38

1 Introduction

In 2009 the Environmental Security Technology Certification Program (ESTCP) sponsored a demonstration of state-of-the-art sensing technologies and analysis routines to characterize subsurface anomalies on munitions ranges. The objective of this demonstration was to explore the anomaly classification performance of the technologies to distinguish unexploded ordnance (UXO) or any items similar to UXO (targets of interest [TOI]) from other items. This demonstration was held at the former Camp San Luis Obispo (SLO) in California. The SLO site is the second such demonstration conducted by ESTCP and was selected for this study due to the challenging terrain which included hillsides characteristic of historical mortar targets[2].

Seven different sensing platforms were employed and multiple classifications algorithms were applied to each sensing platform's data. All classified anomalies on the site were eventually dug up to determine the accuracy of each of the evaluated sensors/classification combinations. The main classification goal of the study was to enable a decision as to which items can be safely left in the ground (i.e. items which are not suspected to be TOI).

The demonstrators were asked to take the information gained from the sensing technologies and create a ranked dig list – the rankings indicated which anomalies were suspected to be clutter (Not-TOI) and hence should not be recovered first during remediation. In many cases these rankings were scores between 0 and 1, where numbers closer to one represent a belief that the anomaly is likely clutter. The common practice for the demonstrators was to select a dig stopping threshold based on the ranking scale metric – and to recommend digging of all anomalies with a ranking metric less than this threshold. Each demonstrator's sensor/classification-efficiency was scored via a receiver operating characteristic (ROC) curve [2].

In total 54 dig lists, which represented combinations of sensor data collections systems, data analysis processing approaches, and disparate demonstrators, were scored in the SLO demonstration. The conclusion drawn from the demonstration was that while classification of subsurface anomalies on the SLO site was more challenging than at the previous demonstration site (Camp Siebert [3]), significant classification ability was indeed demonstrated. Specifically, it was observed that the best performers correctly classified all or nearly all of the targets of interest while achieving reductions of up to 50% in the number of non-TOI anomalies recovered [2].

The demonstration and success of these technologies is indeed encouraging and inspires confidence that they may be applied to munitions range remediation. However, there remains a lack of methods and tools for ongoing monitoring and adjusting as the digging proceeds. We have developed a Bayesian digstop (BDS) methodology which provides in-progress monitoring during the remediation and quantification of the confidence that no TOI remain on site upon completion of the remediation.

We achieve this by building a discrimination methodology around the ranking statistic and can incorporate all available information as anomalies are recovered during remediation. The driver for this work was that, in a real remediation situation, all anomalies are not recovered so an independent measure of remediation progress and resulting confidence attained is desirable [1].

This report documents the application of the BDS algorithm to 25 of the 54 dig lists that were garnered from the SLO demonstration.

2 Bayesian Digstop Methodology

The Bayesian Digstop (BDS) methodology is a statistical algorithm which combines expert judgment, site history and usage data, and data collected from on site undug anomalies to monitor and assess UXO remediation progress on munitions testing ranges. The intent of the algorithm is to provide a quantitative measure of confidence that the remaining undug anomalies are not TOI. TOI's are defined as UXO or anomalies which exhibit UXO characteristics [1].

The BDS algorithm is initialized with any available opinions, history, or data regarding site usage and number of TOI that may be among the undug anomalies. After the initialization step, site remediation commences and after some time BDS can assess remediation progress - the desired result being either a measure of confidence that indicates that remediation thus far has recovered all TOI or an indication that we are not yet confident that all TOI have been recovered. If the latter, this indicates remediation would continue and BDS may again assess remediation progress given the new information gained from digging [1].

We present a flow chart that describes site remediation and monitoring by the BDS algorithm in Figure 1. The major elements of the BDS algorithm are indicated in the grey box.

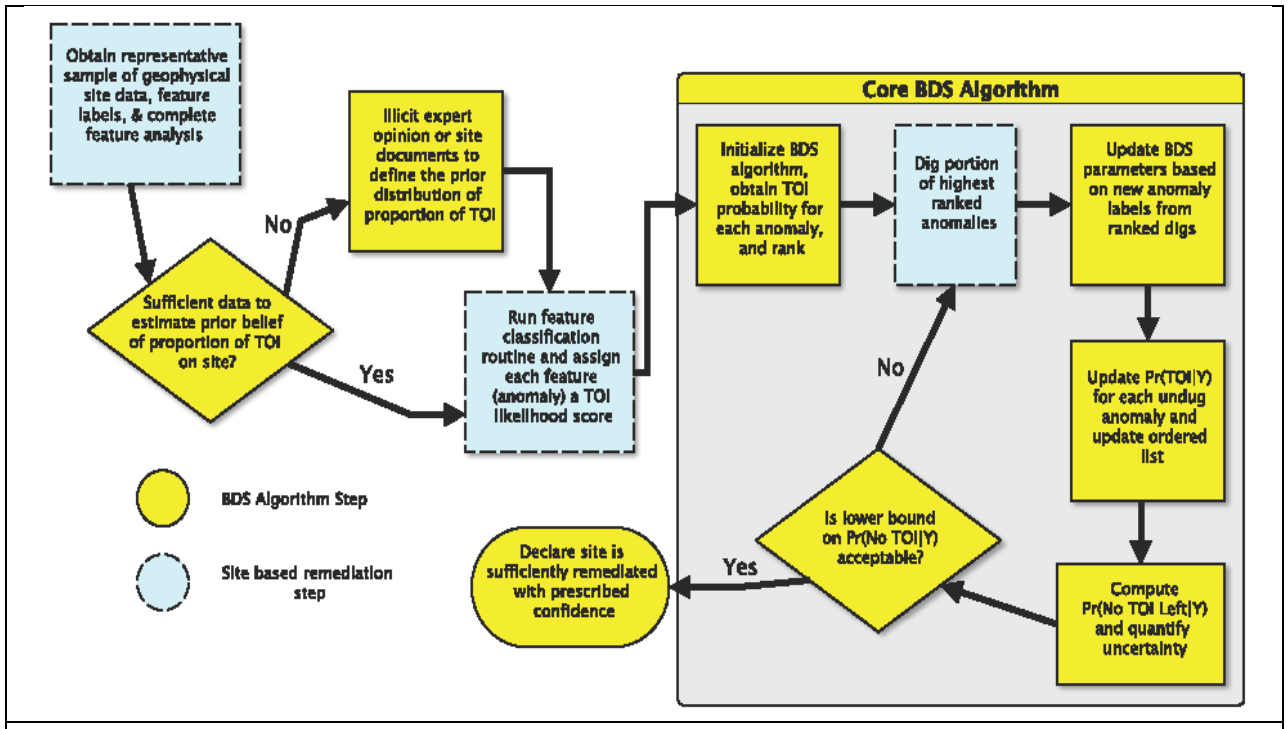

Figure 1: Flow chart depicting the major components of the BDS algorithm as applied to site remediation.

2.1 BDS Objectives and Use

The main objectives of the Bayesian Dig-Stop methodology are as follows.

1. Provide justification for decision to stop digging anomalies during remediation process
2. Quantify the degree of confidence that few or no Targets of Interest (TOI) remain undug
3. Account for past discriminating performance of the sensor/data extraction system using training or historical data
4. Allow updates to the stop digging decision rule and confidence statements as digging proceeds

We have achieved these objectives through the development of a rigorous statistical model which takes advantage of the ranking metric produced by each demonstrator. The ranking statistic represents the degree of belief that the demonstrator has in each anomaly being (Not) TOI. We model the ranking statistic from both the training and available site data to measure the probability that no TOI remain on site at any dig stage. The mathematics of the algorithm can be found in [1]. The algorithm yields two quantities which are used to assess remediation progress:

1. The probability that no TOI remain on site conditional on the observed data, denoted $\Pr(\text{No-TOI Remain} \mid \text{All dug anomaly information})$ and
2. A 95% lower confidence bound (LB) on the quantity in 1.

Specifically, if we observe a sufficiently large 95% LB on $\Pr(\text{No-TOI Remain} \mid \text{All dug anomaly information})$, we infer that all available information indicates that there is a low probability of leaving TOI in the ground and we would conclude a successful site remediation.

2.2 BDS Contributions to the State of the Art

We view the distinguishing characteristics and unique contributions of the BDS algorithm as follows.

1. **Objective statistical confidence support for dig stopping rule** – Our methods rely neither on a single estimate of the probability of TOI remaining on a site nor on a subjective judgment from an analyst on when to stop digging. They recommend stopping only after sufficient confidence is demonstrated that the likelihood of UXO remaining is acceptably small. Thus, when the classification ability of the expert-sensor system is not good (e.g., magnetometer data alone), then to achieve the confidence required, the BDS requires most if not all of the anomalies to be dug. However, when the classification ability is good (e.g., MetalMapper, ESTCP Project MM-0603), confidence can be achieved with fewer digs. *The ability to put the dig-stop decision in the form of a statistical confidence statement is unique and provides a replicable process that can be used and understood by non-experts without as much subjectivity. Further, it takes advantage of the uncertainty in the distribution of TOI metrics through the application of a statistical model to construct the confidence statement*

2. **Explicitly accounting for the training classification performance** – We carry forward valuable information about the discrimination potential given the uncertainties in the metrics for the TOI from the training data. The training data represent the sensor systems performance on a set of anomalies that are assumed to represent the population of undug anomalies. Thus, if the classification was mediocre on the training data set, it will take more site digs to overcome that and achieve an acceptable level of confidence if this is possible. Whereas, if the classification was excellent during training with no TOIs misclassified, then that increases our confidence up front and we will not need as many digs to achieve the confidence levels desired for dig stopping. The BDS methodology quantifies these issues.

3. **Accounting for rarity of UXO** – We can specify a prior based on the observed proportion of TOI on the training data set, on an initial round of semi-random digs from the site, or on expert opinion. As we proceed through digging, this prior can be updated. We can control how much weighting that parameter has so as not to give too much credit for the rarity factor. This parameter may also be used to produce conservative confidence statements.

4. **Standardized method for comparing dig-stopping rules** – During past classification demonstrations, many of the dig stopping rules employed by the performers were highly subjective and inconsistent across the performers. As it currently stands, the BDS methodology can be viewed as a "standardization" of the dig stopping decisions made

by disparate demonstrators and sensing systems. The confidence statement can be used to quantify and compare the different dig stopping rules made by different demonstrators.

5. **Accounts for distribution of all anomalies metrics** – instead of using the ranking statistics and the underlying feature results on an anomaly by anomaly basis, we examine and model the entire distributions of the anomaly statistics for the TOI and Not-TOI populations. This enables a comparison of the two populations.

6. **An independent analysis of the discriminatory ability of the ranking metrics** – the ranking metrics are produced by each demonstrator after physics-based and statistical analyses of the anomaly data collected by the sensors. The intention of the ranking metric was to provide field workers with a prioritized ordering of sensor anomalies which was used as a guide to which anomalies should be recovered first. Since BDS is at its heart a discrimination approach it also indicates how well the ranking metrics actually order and prioritize the digging of the anomalies.

3 BDS Application to SLO Data

Of the 54 dig lists that were produced during the SLO demonstration, we have analyzed 25 of them with the BDS algorithm. The sensor system/demonstrators considered in this report are indicated in Table 2. The names in Table 2 were gleaned from the SLO data base [2]. In the remainder of this document we will refer to a particular dig list by combining the names across the rows of Table 2. Since the volume of data for this demonstration was much greater than the Camp Siebert demonstrations [3], and since this demonstration was conducted over more difficult terrain, in addition to the BDS analysis we've also implemented simple model diagnostics. The diagnostics produce a quantification of how well the assumptions upon which BDS is built hold in the data – conclusions made from the BDS algorithm are considered valid if the modeling assumptions hold. The largest assumption imposed by the BDS model is that the ranking metrics for both the TOI and Not-TOI populations can be modeled with a *Beta* distribution. We will inspect this assumption through the use of a Beta quantile-quantile (QQ) plot.

Demonstrator/ Feature Extraction	Sensor	Mode
SIG	EM61 Array	PNBC
SIG	EM61 Array	RVM
SIG	MAG	PNBC
SIG	MAG	RVM
SIG	MSEMS	PNBC
SIG	TEMTADS	PNBC
SIG/UBC	EM61 Array	PNBC
SIG/UBC	EM61 Array	RVM
SIG/UBC	EM61 Cart	PNBC
SIG/UBC	MAG	PNBC
SIG/UBC	MAG	RVM
SIG/UBC	MSEMS	PNBC
SIG/UBC	TEMTADS	PNBC
SIG/UBC	TEMTADS	RVM
SAIC/UXAnalyze	EM61 Array	
SAIC/UXAnalyze	EM61 Cart	No Slope Correction
SAIC/UXAnalyze	EM61 Cart	Slope Correction
SAIC/UXAnalyze	MSEMS	
SAIC	TEMTADS	2 Criteria
SAIC	TEMTADS	3 Criteria
SAIC	MetalMapper	2 Criteria
SAIC	MetalMapper	3 Criteria
Sky	Metal Mapper	
SIG	Metal Mapper	PNBC
SIG	MetalMapper	RVM

Table 2: The 25 dig lists considered in this analysis as indicated by Demonstrator, Sensor, and Mode.

As an example of the analysis results, we present a panel of graphics for the SAIC_MetalMapper_3criteria dig list in Figure 2. We now describe the suite of graphics and their interpretation. In the top left panel we present a graphic of the BDS performance, as measured by the Pr(No TOI's Remain | all dug data) (black line) and its 95% lower confidence bound (red dotted line) as we move the dig threshold through the ranking statistic (right to left). Thus this graphic indicates that, for the SAIC_MetalMapper_3criteria dig list, a dig threshold around 0.65 yields large probability that No TOI remain on site with a large 95% confidence bound. In the top right panel, we present a ROC performance curve which plots the Proportion of TOI Discovered vs. the Number of Anomalies dug. Indicated on this plot in the green and red points are the 95% lower bound (0.95 and 0.99 respectively), the dig threshold that achieved these lower bounds (0.68 and 0.60 respectively), and the number of anomalies dug at this threshold. This ROC curve indicates that 5 and 3 TOI's were not recovered at these dig thresholds respectively. The middle two plots show the histograms of the ranking metric for all of the TOI and Not-TOI anomalies; the red curve on each plot indicates the fitted Beta distribution – these plots give an initial

inspection to how well a Beta distribution can approximate the densities of the ranking metric. The bottom two plots are the Beta QQ-plots for the TOI and Not-TOI density fits respectively. These graphics plot the sample quantiles of the ranking metric on the vertical axis vs. the theoretical quantiles from the fitted Beta distribution on the horizontal axis – thus enabling an inspection of how well the Beta distribution fits the observed ranking metrics. If the points fall on the blue dotted line this provides some evidence that the ranking metric can be approximated by the Beta distribution. The red dotted lines are a 95% confidence interval for the quantiles – if the points fall wholly within these bands then we can conclude that the Beta fit is adequate. Thus, for the TOI fit, we observe all of the points on or within the bands – indicated that the Beta fitted to the TOI ranking metrics is somewhat reasonable. In contrast the points on the Not-TOI QQ-plot deviate from the blue line and out of the red bands in the middle of the (0,1) range – this calls into question somewhat the Beta fit, but we note that the deviation is not too severe.

To summarize the story told by this suite of graphics, we observe that, for the SAIC MetalMapper dig list, BDS would provide high confidence that no TOI remain on site early in the remediation process (dig threshold of 0.60 on the ranking statistic) saving a significant number of anomaly digs, however, 3 TOI would be left in the ground. These results show an interesting phenomenon with the SAIC process. Their techniques do quite well at discriminating between TOI and scrap. In fact, their method is one of the better performers in terms of items left in the ground. The three items left in the ground may be more of a data processing problem instead of an algorithmic problem.

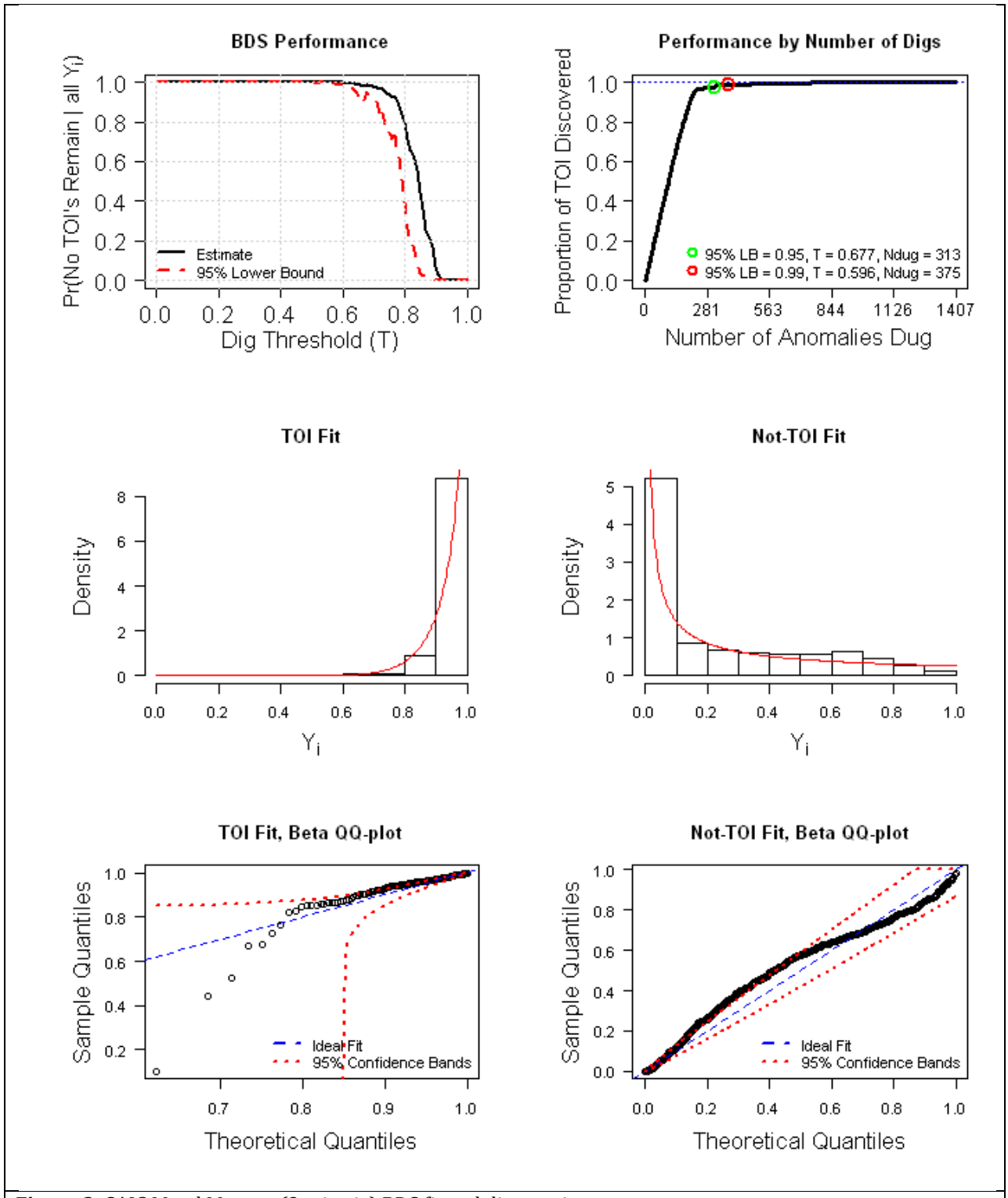

Figure 2: SAIC Metal Mapper (3 criteria) BDS fit and diagnostics.

As another example we present the performance and diagnostic graphics for the SIG-UBC TEMTADS (RVM) sensor/demonstrator in Figure 3. The BDS Performance curve indicates that a dig threshold of 0, which translates into digging almost all anomalies, is required to achieve reasonable confidence that no TOI remain on site. Similar conclusions are also indicated in the ROC curve. Inspection of the histograms of the TOI and Not-TOI ranking metric and fitted densities gives an initial impression that the Beta fits are not reasonable. We do note, however, that Not-TOI ranking statistic is dominated by values close to 0 – an artifact that is difficult to incorporate into a Beta model. The QQ-plot for the TOI density indicates reasonable fit of the Beta to the ranking metric distribution. The QQ-plot for the Not-TOI density indicates some deviation of fit in the early quantiles (due to the majority of metric values being 0) but the overall fit of the Beta distribution seems to capture the properties of the ranking metric for the Not-TOI's. The overall story told by the plot is that application of the BDS model is reasonable for these data. The BDS calculation for this technology/classification combination requires recovery of all anomalies to produce an appropriate lower confidence limit on the probability that no TOI remain which indicates that the site is clean. This conclusion seems to be the correct conclusion given that there are indeed TOI's whose metric are near 0, thus requiring a low dig threshold to recover them.

We present the performance and diagnostic graphics for the remaining sensor/demonstrators in the appendix without discussion.

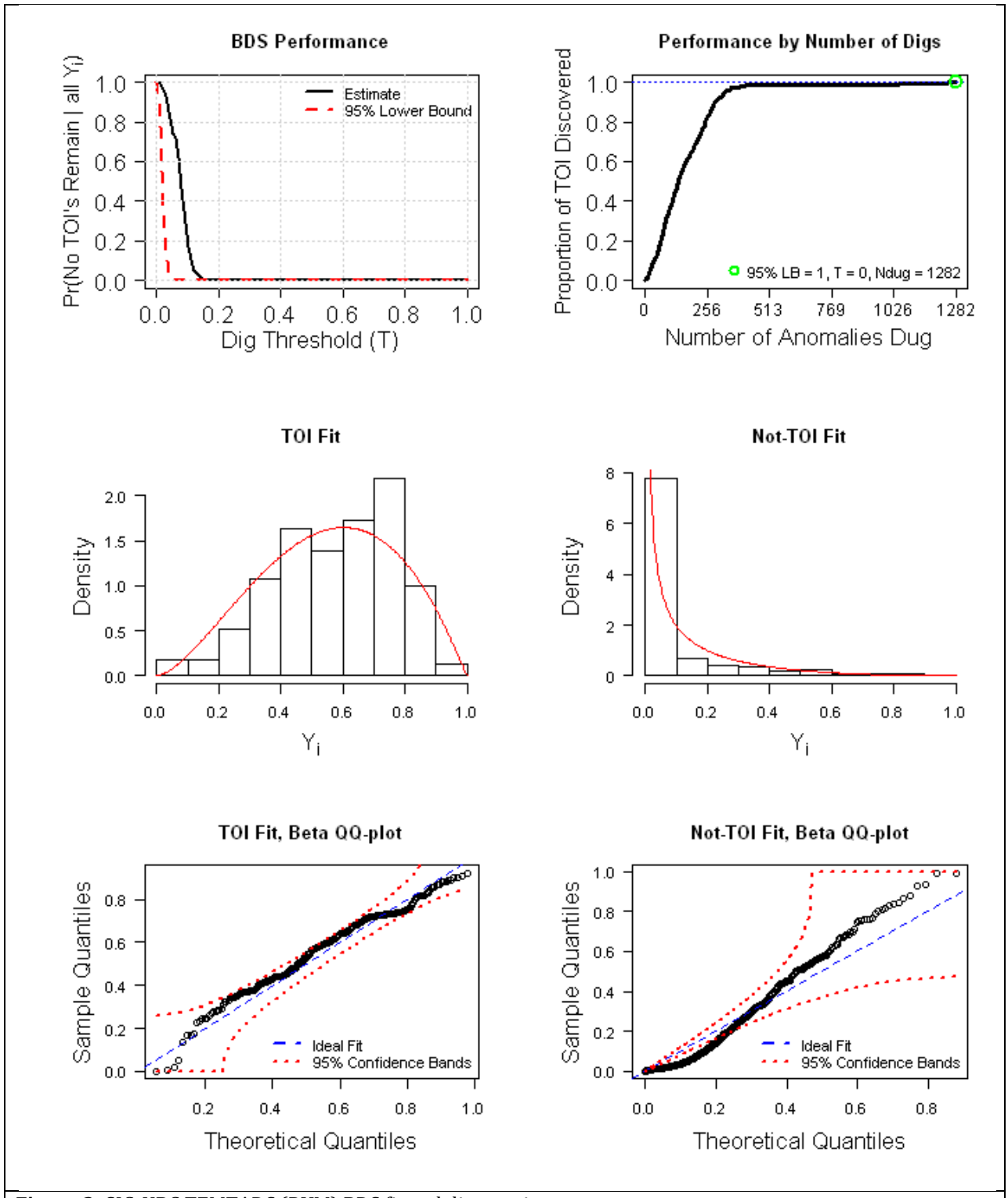

Figure 3: SIG-UBC TEMTADS (RVM) BDS fit and diagnostics.

3.1 Results Summary

Table 3 presents a summary of the application of BDS to the 25 SLO dig lists; we report

1. A 95% lower confidence bound on the probability that no TOI remain; that would likely lead to conclusion of digging
2. The number of anomalies dug under the conclusion
3. The number of TOI that remain unrecovered
4. An indication of whether the Beta fits to the TOI and Not-TOI ranking statistics were reasonable.

The rows highlighted in green indicate sensor/demonstrator's dig lists where BDS indicated high confidence that no TOI remain, where a significant number of anomalies were not recovered, and among those anomalies not recovered either 0 or 1 TOI remained. Among these the SAIC/UXAnalyze_EM61_Array, SAIC/UXAnalyze_EM61_Cart_noslopecorrection, SAIC/UXAnalyze_EM61_Cart_slopecorrection, and SAIC/UXAnalyze_MSEMS sensor/demonstrators exhibited a dig-stop conclusion which saved a significant number of anomaly digs with high confidence that no TOI remain. The distributions of the TOI and Not-TOI ranking metrics were reasonably well characterized by the fitted Beta distributions, and only 1 TOI was not recovered. We note that the distributions of the ranking metric for the TOI and Not-TOI anomalies were reasonably well separated- this enables good discrimination ability. We also note that the single unrecovered TOI for these sensor/demonstrators did not fit the overall pattern of the ranking metric distribution for the TOI's; they had a ranking metric near 0 and would have required digging all anomalies in order to recover them. The SIG_MetalMapper_PNBC and SIG_MetalMapper_RVM sensor demonstrators resulted in a dig stop conclusion which saved a significant number of digs, yielded high confidence that no TOI remain, and indeed recovered all TOI anomalies. However, the diagnostics indicate that the distribution of the TOI and Not-TOI metrics were not well characterized by the Beta distribution – the reason being that some of the distribution were bimodal, a characteristic that is not within the ability of the Beta distribution to model. In practice, this would call into question our "confidence" in the claims made by the BDS algorithm. We note that the observed "good performance" of BDS on these two sensor/demonstrators is largely due to the fact that the modes of the TOI and Not-TOI ranking metric distributions are far from each other.

The rows highlighted in yellow indicate sensor/demonstrator's dig lists where BDS indicated high confidence that no TOI remain, where a significant number of anomalies were not recovered, but where there were 2 or more TOI which remained unrecovered. Most of the distributions of TOI and Not-TOI ranking metrics were fairly well characterized by the Beta fit for these demonstrators. As a general observation, many of these dig lists contained TOIs whose ranking metric was low and well within the high density region of the Not-TOI distribution – BDS could not detect these TOIs. The results for the SAIC_MetalMapper_3crit are worth commenting on. The distributions of the TOI and Not-TOI metrics were well separated, in fact, so much so that the BDS algorithm basically determined it was finished before removing the last few TOI.

	All columns use the 95% lower confidence bound					
	Number of anomalies undug		Proportion of anomalies undug		Number of TOI remaining	
Sensor/Demonstrator	95% probability that no TOI remain	99% probability that no TOI remain	95% probability that no TOI remain	99% probability that no TOI remain	95% probability that no TOI remain	99% probability that no TOI remain
SIG_MetalMapper_PNBC	928	647	0.65	0.46	0	0
SAIC/UXAnalyze_EM61 Array	292	204	0.22	16	1	0
SAIC/UXAnalyze_EM61 Cart No slope correction	604	592	0.47	0.46	1	1
SAIC/UXAnalyze_EM61 Cart Slope correction	605	582	0.47	0.45	1	1
SAIC/UXAnalyze_MSEMS	455	445	0.35	0.34	1	1
SIG_MetalMapper_RVM	1163	1049	0.82	0.74	7	2
SAIC_TEMTADS_2 Criteria	851	823	0.68	0.64	3	3
SAIC_TEMTADS_3 Criteria	650	547	0.51	0.42	3	3
SAIC_MetalMapper_2Criteria	1070	1052	0.76	0.74	4	3
SAIC_MetalMapper_3Criteria	1060	954	0.75	0.68	3	2
Sky_MetalMapper	1054	993	0.75	0.71	2	2
SIG/UBC_MAG_PNBC	350	205	0.30	0.18	7	2
SIG_EM61 Array_RVM	< 271	NA	< 0.24	NA	3	NA
SIG_MAG_PNBC	NA	NA	NA	0	0	0
SIG/UBC_EM61 Cart_PNBC	NA	NA	NA	0	0	0
SIG/UBC_MAG_RVM	NA	NA	NA	0	0	0
SIG/UBC_MSEMS_PNBC	NA	NA	NA	0	0	0
SIG/UBC_TEMTADS_PNBC	< 131	NA	< 0.10	NA	1	NA
SIG/UBC_TEMTADS_RVM	< 426	NA	< 0.33	NA	2	NA
SIG_EM61 Array_PNBC	42	NA	0.04	NA	0	NA
SIG_MAG_RVM	35	31	0.027	0.025	1	1/1
SIG_MSEMS_PNBC	< 204	NA	< 0.16	NA	0	0/0

Table 3: BDS summary of results as applied to the 25 SLO dig lists. Green highlighting indicates dig lists where BDS produced large 95% LB on Pr(No TOI Remaining | Data), a significant number of anomaly digs were saved, and where only 0 or 1 TOI were left unrecovered. Yellow highlights indicate dig lists where a large 95% LB was produced by BDS, where a significant number of anomaly digs were saved, but where 2 or more TOI were left unrecovered. Red highlighting indicates dig lists where BDS was unable to produce high confidence that all TOI were recovered. White highlighting indicates dig lists where BDS measured high confidence that all TOI could be recovered at some dig threshold, but where the threshold was so low that almost all anomalies needed to be recovered.

The rows highlighted in white indicate sensor/demonstrator's dig lists where BDS required digging almost all anomalies to achieve a high confidence that no TOI remain – all of these resulted in 0 or 1 remaining TOI in the unrecovered anomalies. These results can be summarized as follows: first, most of the Beta fits were reasonable for these dig lists, second, the BDS dig stop decision was correct, meaning that we would have to dig almost all anomalies based on the ranking statistic to get all TOI's, last, this is due to either poor separation of the TOI and Not-TOI ranking statistics distributions or very broad TOI distributions with TOI anomaly scores near 0.

Last, the rows highlighted in red indicate sensor/demonstrator's dig lists where BDS was not able to achieve any confidence that no TOI remain on site – this is with recovering all anomalies. Generally, these conclusions are the result of the TOI and Not-TOI metric distributions being almost identical (no discriminatory ability) for these demonstrators – the conclusion is correct.

To summarize at a high level, regardless of the appropriateness of the Beta fits as indicated by the diagnostics, the dig lists highlighted in green demonstrate ranking statistics whose distributions for the TOI and Not-TOI populations were reasonably well behaved and separated – this indicates that the sensing and classification system which produced these dig lists may produce reliable conclusions in the future. The dig lists highlighted in yellow were often plagued by TOI's who were assigned a ranking statistic near 0 – outliers that did not fall within the distribution of most TOI's. The dig lists highlighted in red were plagued by poor discrimination – the TOI and Not-TOI distributions of the ranking statistics were either identical or the TOI distribution was nested well within the Not-TOI distribution. Any automated algorithm would find it difficult to discriminate in this setting. Last, the dig lists highlighted in white can be generally characterized as some discrimination in the ranking statistic among the TOI and Not-TOI populations, but the TOI distribution often had large density near 0 which implicated that near all anomalies needed to be recovered.

4 Conclusions and Recommendations

In this body of work we laid the initial foundation to producing on line diagnostics for a 'real life' TOI remediation scenario. The initial look into diagnostics as applied to the SLO dig lists indicate that some of the ranking statistics distribution were not well characterized by fitted Beta distributions – an observation related to the fact that the SLO demonstration incorporated much more challenging geography than the previous Siebert demonstration.

Since the BDS algorithm currently employs the anomaly ranking statistics, the result of application of the BDS algorithm to the dig lists will, in part, describe how well the ranking statistics are actually able to discriminate between the TOI and Not-TOI anomalies. Thus, we were able to 'rank' the dig lists in regard to their ability to remediate near all TOI and save a number of anomaly digs. Table 1 indicates this ranking (by color category). The quantification of the dig lists ability lies in the 95% LB on Pr(No TOI remain | Data) – this enables cross comparison of the dig lists.

We note that without the ability to evaluate the dig list performance via a ROC curve, the BDS performance plot may be produced from data on available dug anomalies. The intent of the statistical diagnostics is to provide an added level of confidence that BDS is appropriate for the observed data.

Several issues have been identified and need to be resolved before this BDS methodology is deemed defensible in general practice and application. It should be noted that these issues are not peculiar to just the BDS methodology. Any dig-stop decision support approach must

also deal with these issues. Additional research and evaluation is required to address these issues. The key issues are listed below followed by proposed research paths to address them.

Issue 1*: Using actual classification results from demonstrators that used site data (Sibert and SLO), BDS dig stopping rules based on conditional confidence statements have often resulted in a few TOI that are not dug. Thus, it appears that claims about the stated confidence of no TOI remaining based on the BDS methodology are not as strong as the data support. This is in large part due to the fact that the training data sets did not include TOI that were similar to those that were missed and partly because the feature characterization may have resulted in a scoring metric that obscured TOI features related to the anomaly*.

Recommended Resolution: A major component of the BDS methodology is the initialization of the algorithm using the training data. It was frequently observed that while the training data did represent the bulk of the distribution of the TOI anomalies in the ranking statistics, it often failed to account for a very few TOIs in the lower tail of the dig set of anomalies. In real application, there is no available information that such anomalies exist on site. Thus, if training data will be used in application, adequate care must be taken to elucidate the character of the entire distribution of the TOI anomalies (this is not a PNNL task).

We have also developed a set of diagnostic tools that elucidate how well the BDS model fit the data. Inferences (claims and decisions) made under any statistical model are *only as valid and trustable as the model adequately represents the observed data*. While it is not possible to create a general methodology that can be generally applied and models all possible data scenarios, it is possible to construct and apply a general methodology which takes advantage of the common characteristics which give rise to the data. If the model fits the given data well, the confidence statements should be accurate and valid. If the model does not fit the data well then we should "trust" less the confidence statements. This goodness of fit information has been presented in Table 2.

Issue 2*. The BDS algorithm employs the ranking statistic (likelihood of being TOI) generated by all classifiers. However, this coarse metric integrates many considerations from the feature space and significant loss of information may result. This loss of information is one hypothesized cause of the poorly performing confidence statements as the uncertainty in fitting features may not be completely represented in the scoring metric. Moreover, there is little consistency in the development of this scoring metric so significant subjectivity may be involved.*

Recommended Resolution: A path to investigate the loss of resolution hypothesis and mitigate the subjectivity introduced by the experts is to move the BDS algorithm closer to the data. It would be good to directly model the anomaly feature data collected by the sensors. This data characterizes the subsurface anomalies through a number of electromagnetic measurements. These measurements produce a fingerprint for each anomaly and the resulting characteristics inform whether each anomaly may be TOI or not. These are the data that are sometimes subjectively, sometimes rigorously interpreted by the expert to produce the ranking statistic.

Since the BDS method may be considered quite general, it is straightforward to conceptualize how the BDS method can be extended to work in this multivariate feature space. The working hypothesis is that the multivariate data may provide features which are useful in discriminating TOIs from Not-TOIs and that these features are currently "marginalized out" in the BDS method through the use of the ranking statistics.

Issue 3. *It is unclear whether seeded training sets that cover the range of possible TOI configurations will be included for each site investigated. Although the BDS methods can start with an uninformed prior and update or train the Bayesian priors as digging proceeds, one should explore how performance data on previous sites can be incorporated.*

Recommended Resolution: This is related to issue 1. If it is believed that past data on a sensing system represents the system performance in a new application, then it may be rational to compile the past data in the hope of getting a more accurate characterization of the distribution of TOI rankings on site. If this is a valid approach, then the initialization of the BDS methodology on a new application will better reflect what is expected to be encountered on site. By using data from past sensor/algorithm training and site applications, a more representative set of data would be available to ensure good coverage of the possible TOI feature space. A library of past sensor/algorithm training and observational data could be developed and evaluated for consistency across site conditions, munitions used, and test configurations. The BDS methodology could then be adapted to incorporate various consistent subsets of data from these historical training and demonstration exercises to train the Bayesian priors based on the actual site information.

References

1. Walsh, S., Anderson, K., Hathaway, J., and B. Pulsipher. Bayesian Dig Stop: A methodology for Monitoring and Assessing the Remediation of UXO from Munitions Testing Ranges. ESTCP: Statistical Verfication and Remediation Sampling Methods (200837). 2009.

2. Nelson, H., Andrews, A., and Kaye, K. Final Report: ESTCP Pilot Program, Classification Approaches in Munitions Response San Luis Obispo, California. 2010.

3. Nelson, H., Kaye, K., and Andrews, A. Final Report: ESTCP Pilot Program, Classification Approaches in Munitions Response. 2008.

Appendix

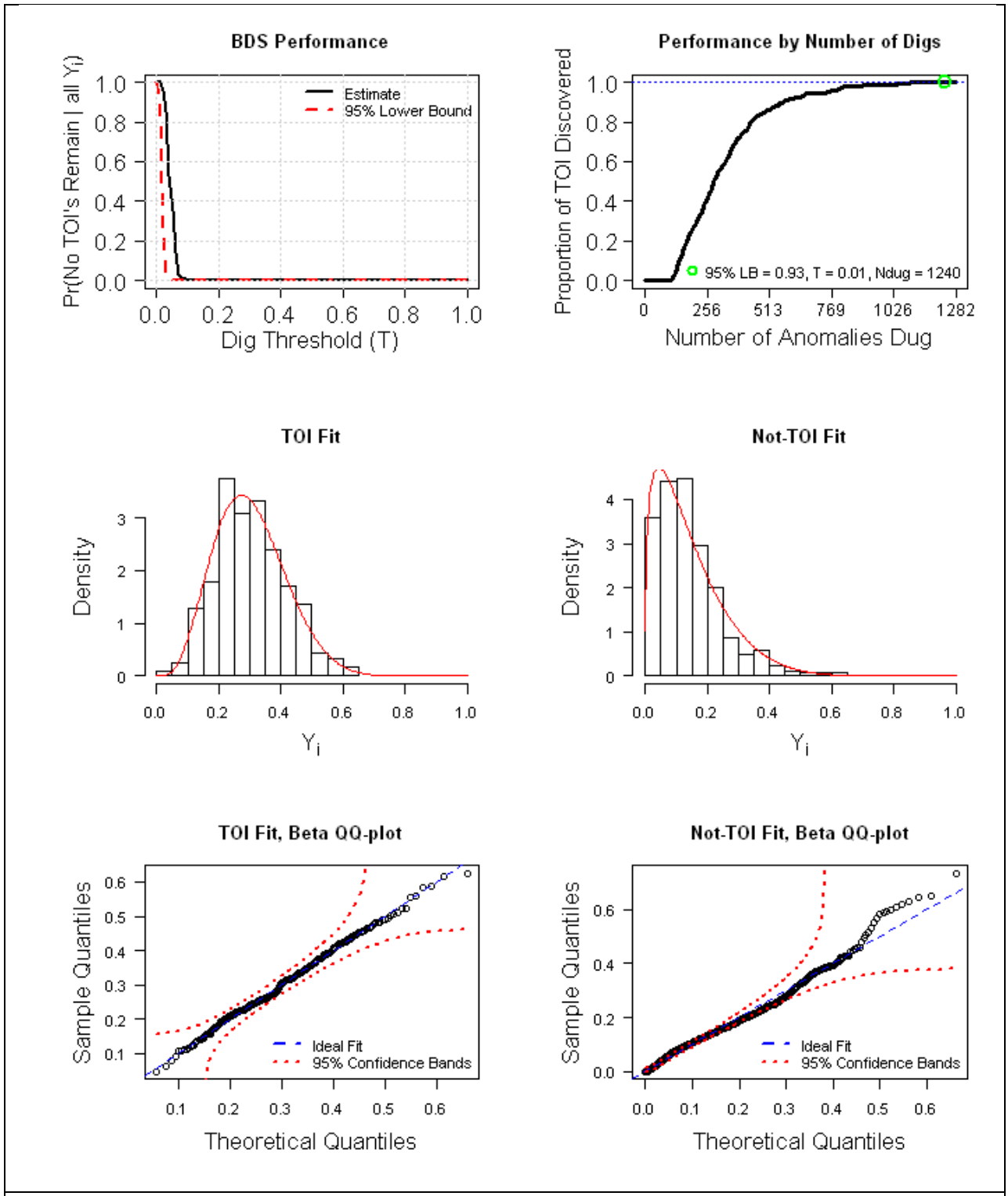

Figure A 1: SIG EM61 Array (PNBC) BDS fit and diagnostics.

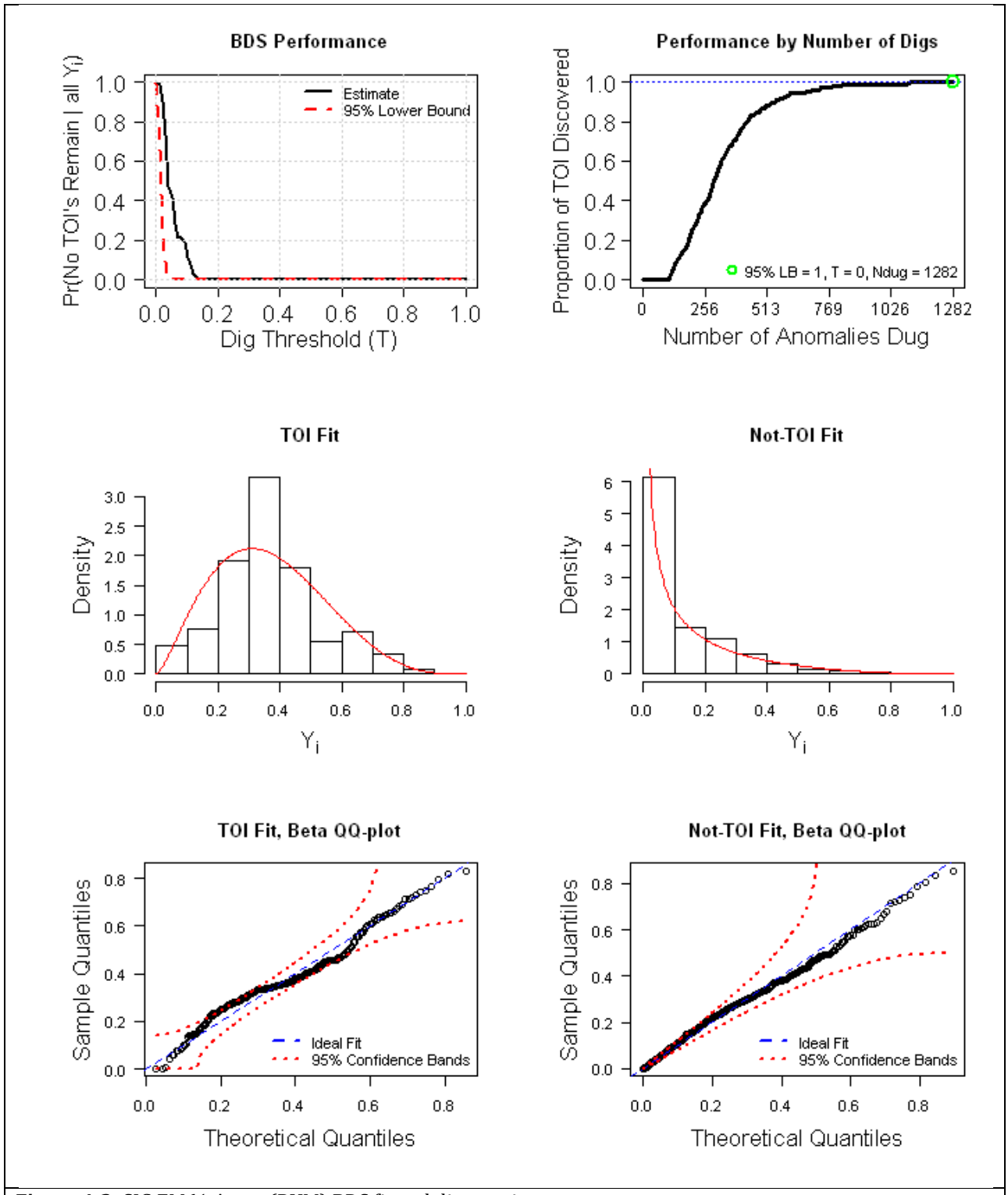

Figure A 2: SIG EM61 Array (RVM) BDS fit and diagnostics.

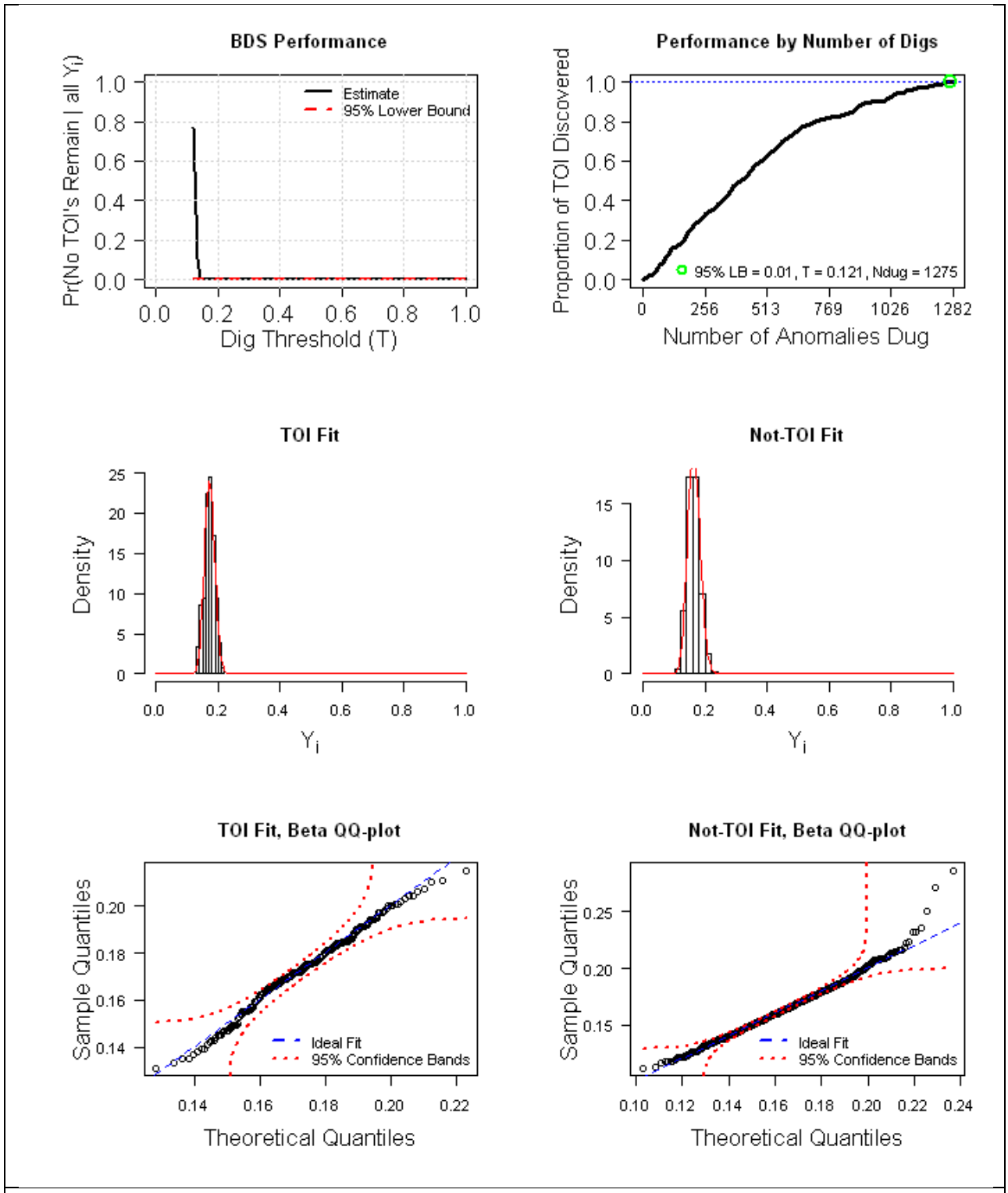

Figure A 3: SIG MAG (PNBC) BDS fit and diagnostics.

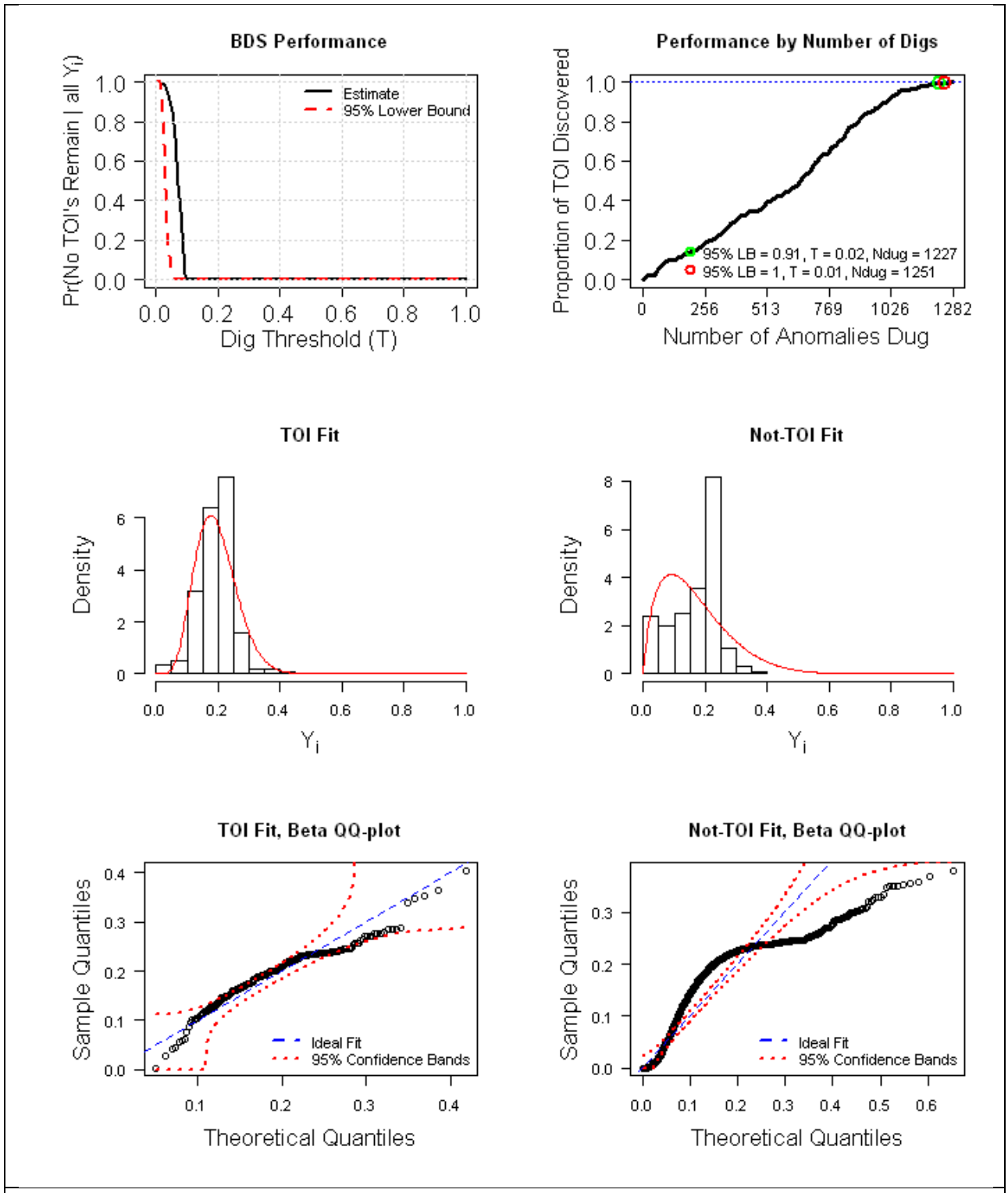

Figure A 4: SIG MAG (RVM) BDS fit and diagnostics.

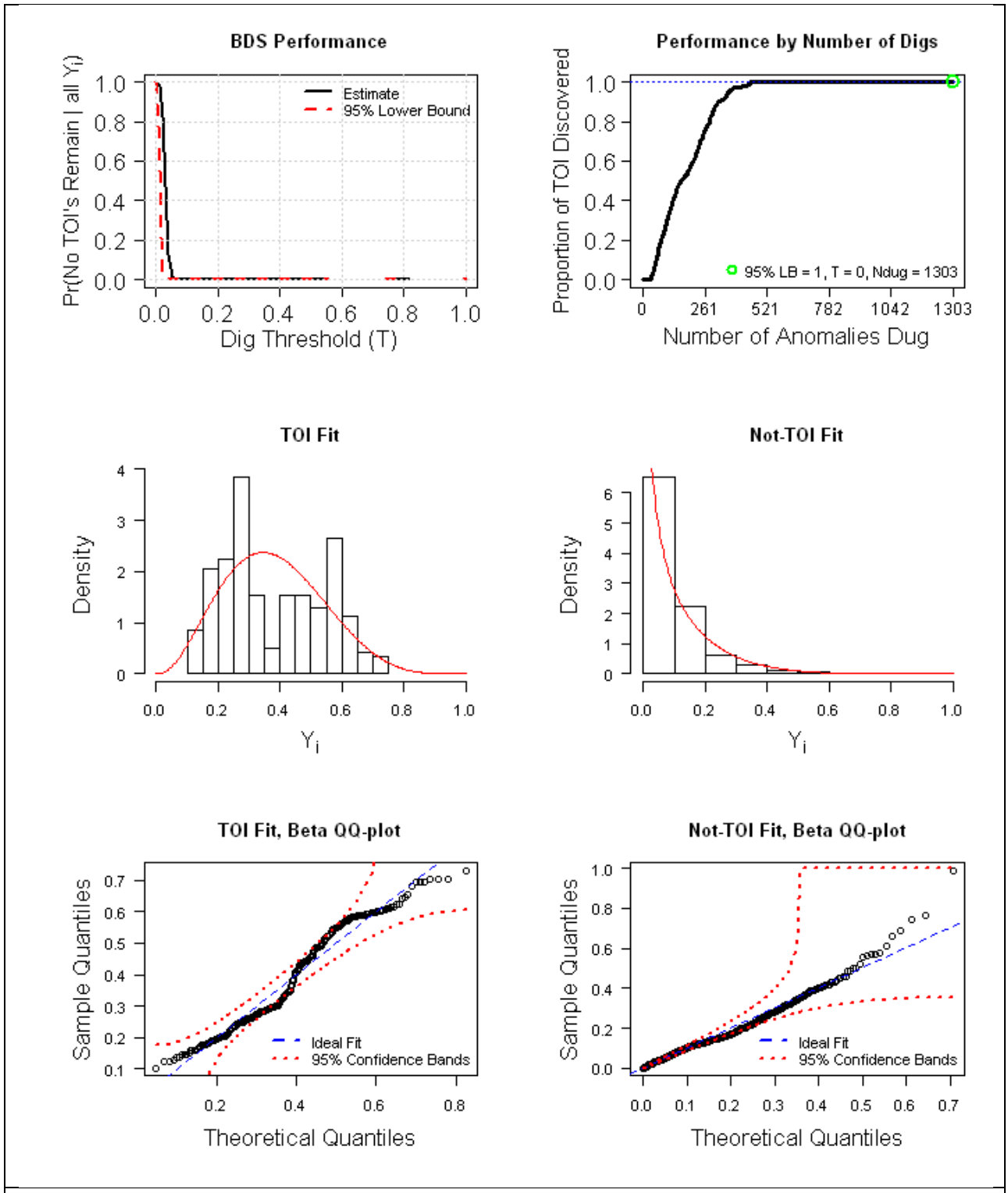

Figure A 5: SIG MSEMS (PNBC) BDS fit and diagnostics.

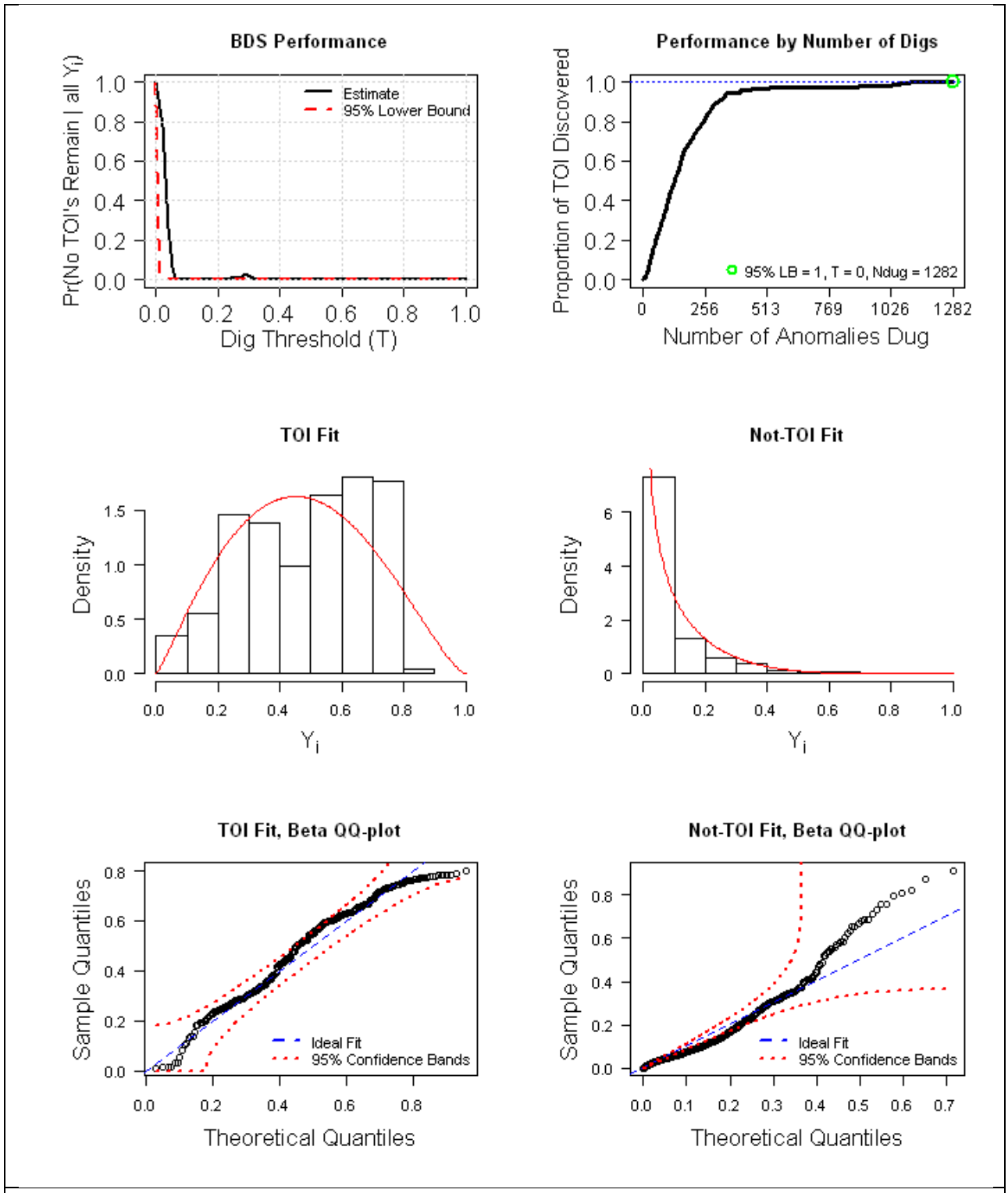

Figure A 6: SIG TEMTADS (PNBC) BDS fit and diagnostics.

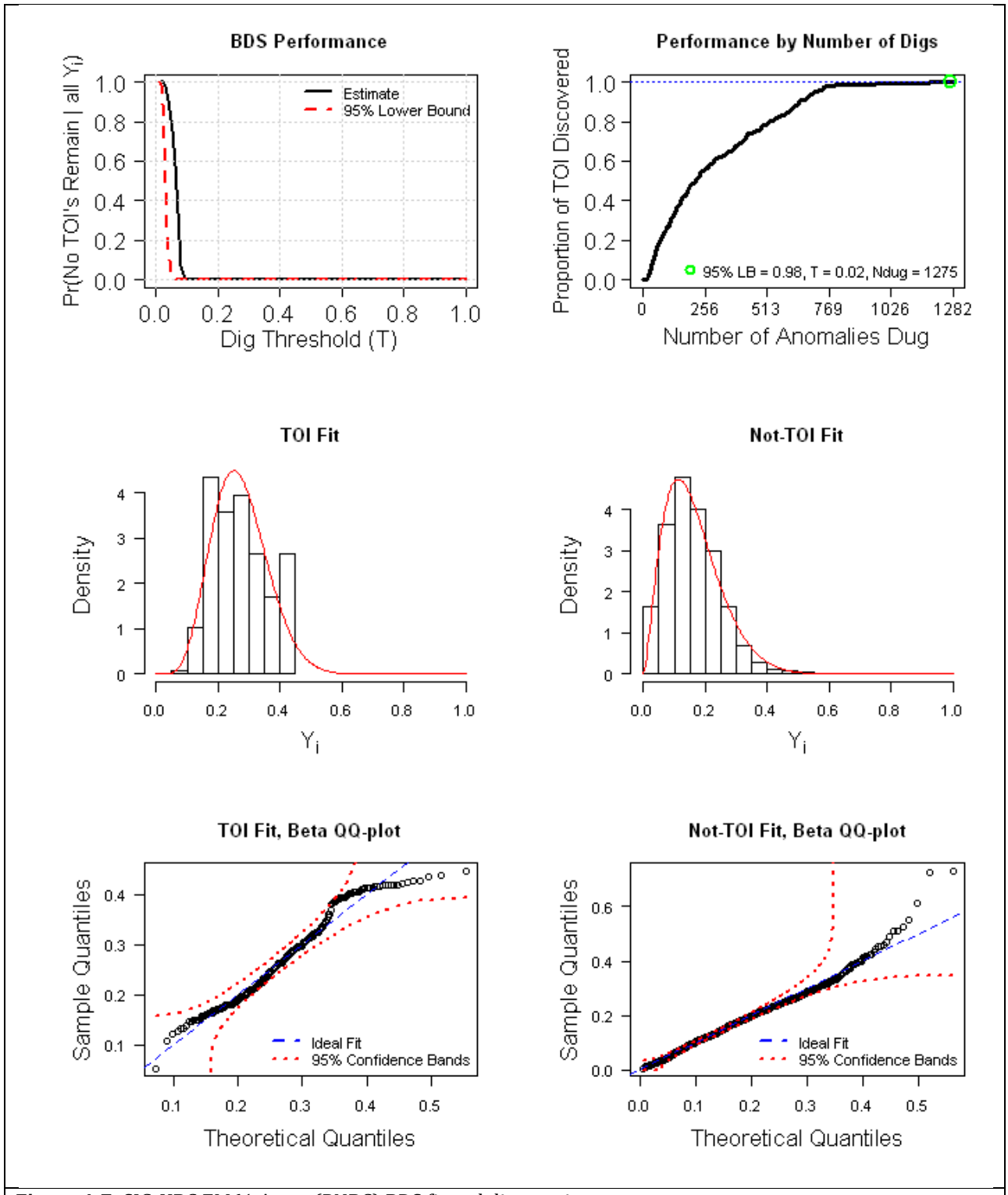

Figure A 7: SIG-UBC EM61 Array (PNBC) BDS fit and diagnostics.

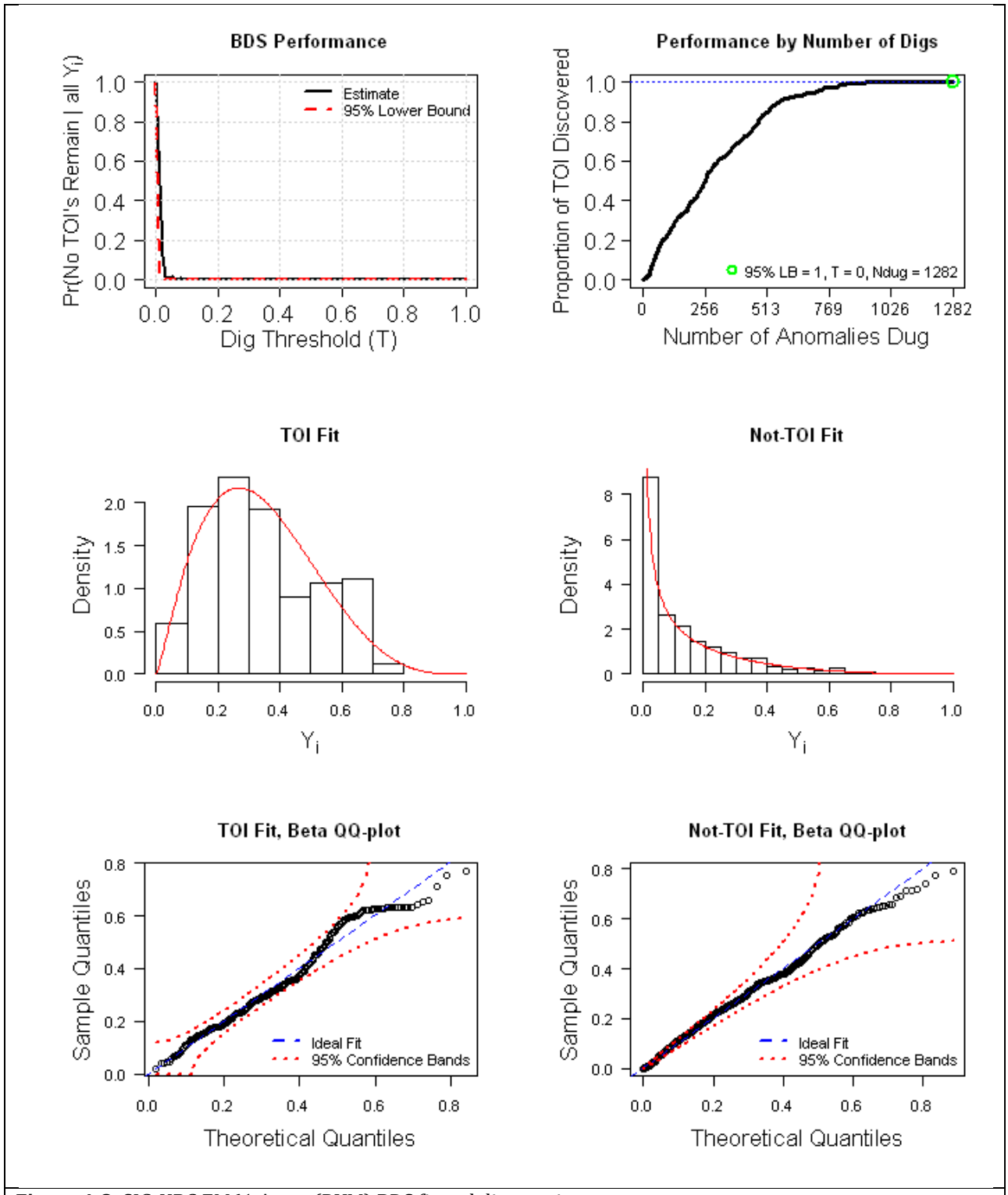

Figure A 8: SIG-UBC EM61 Array (RVM) BDS fit and diagnostics.

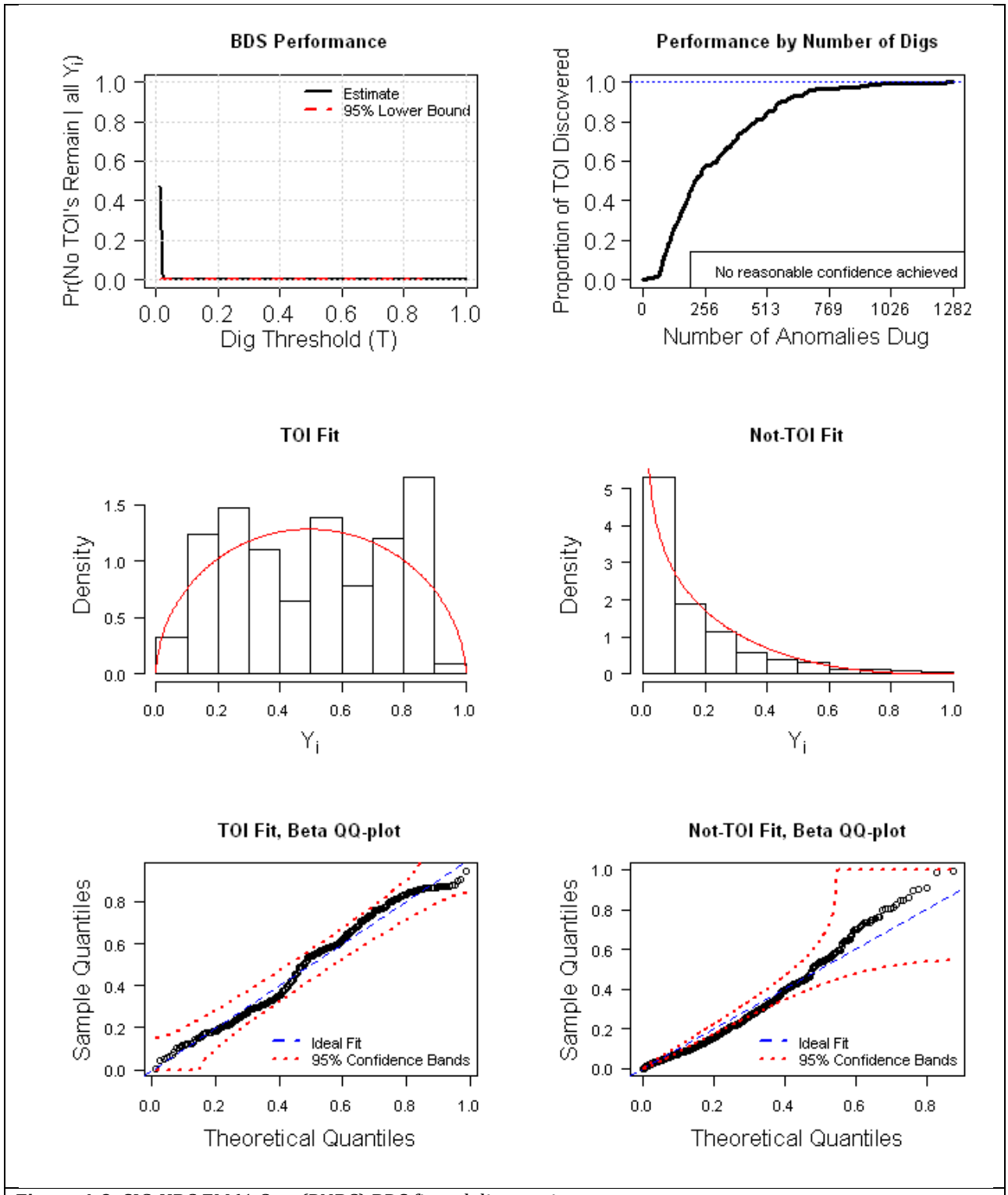

Figure A 9: SIG-UBC EM61 Cart (PNBC) BDS fit and diagnostics.

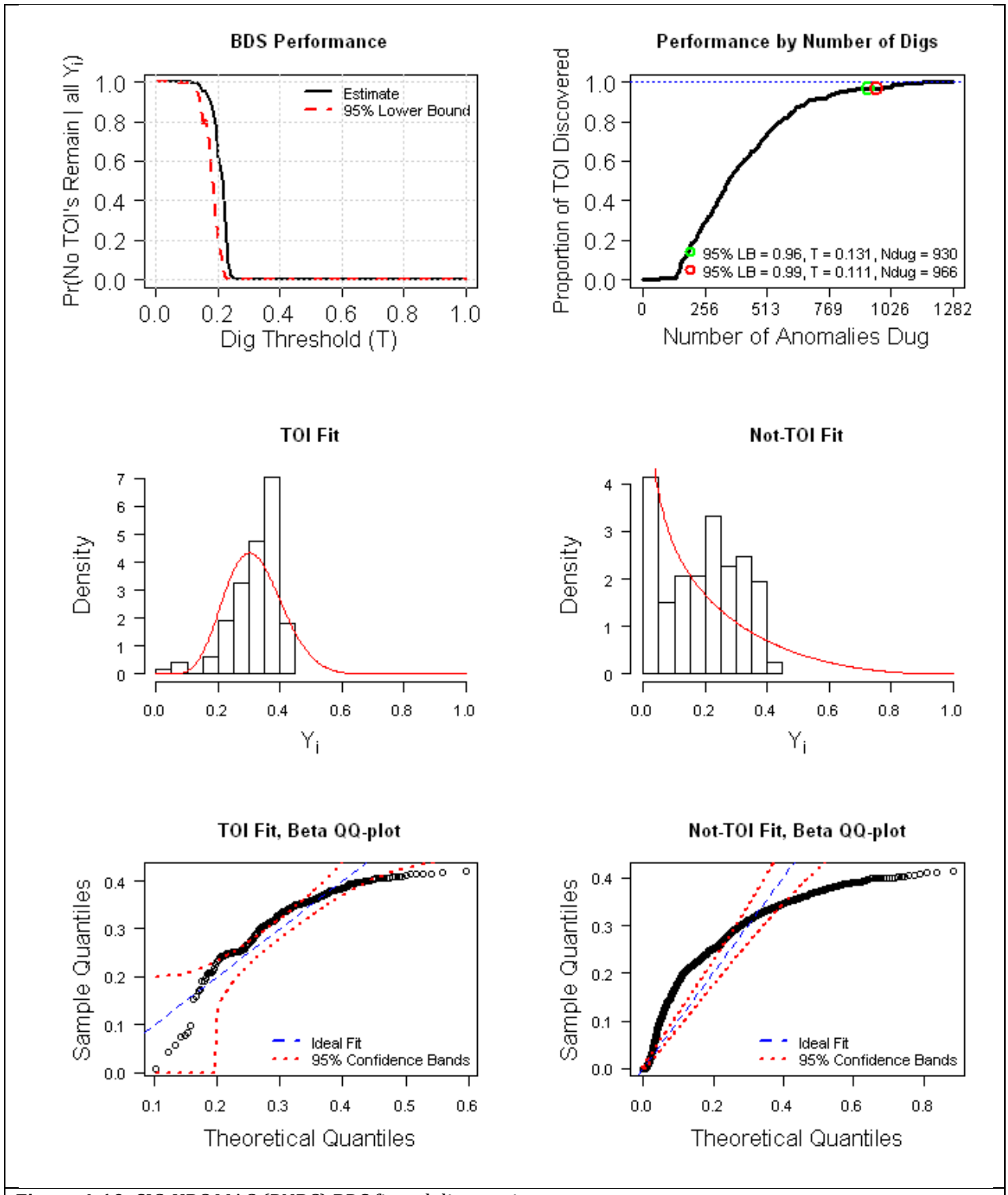

Figure A 10: SIG-UBC MAG (PNBC) BDS fit and diagnostics.

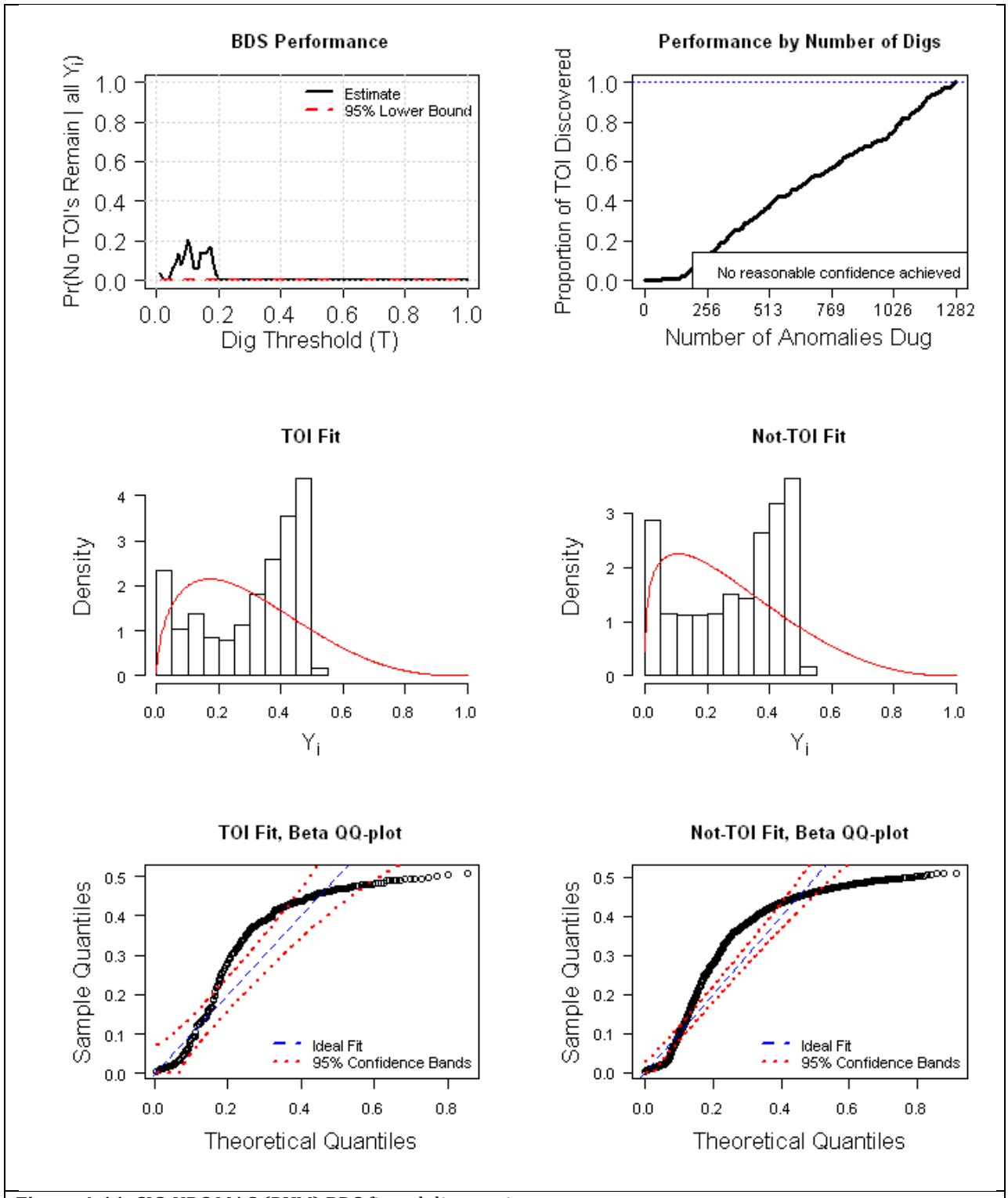

Figure A 11: SIG-UBC MAG (RVM) BDS fit and diagnostics.

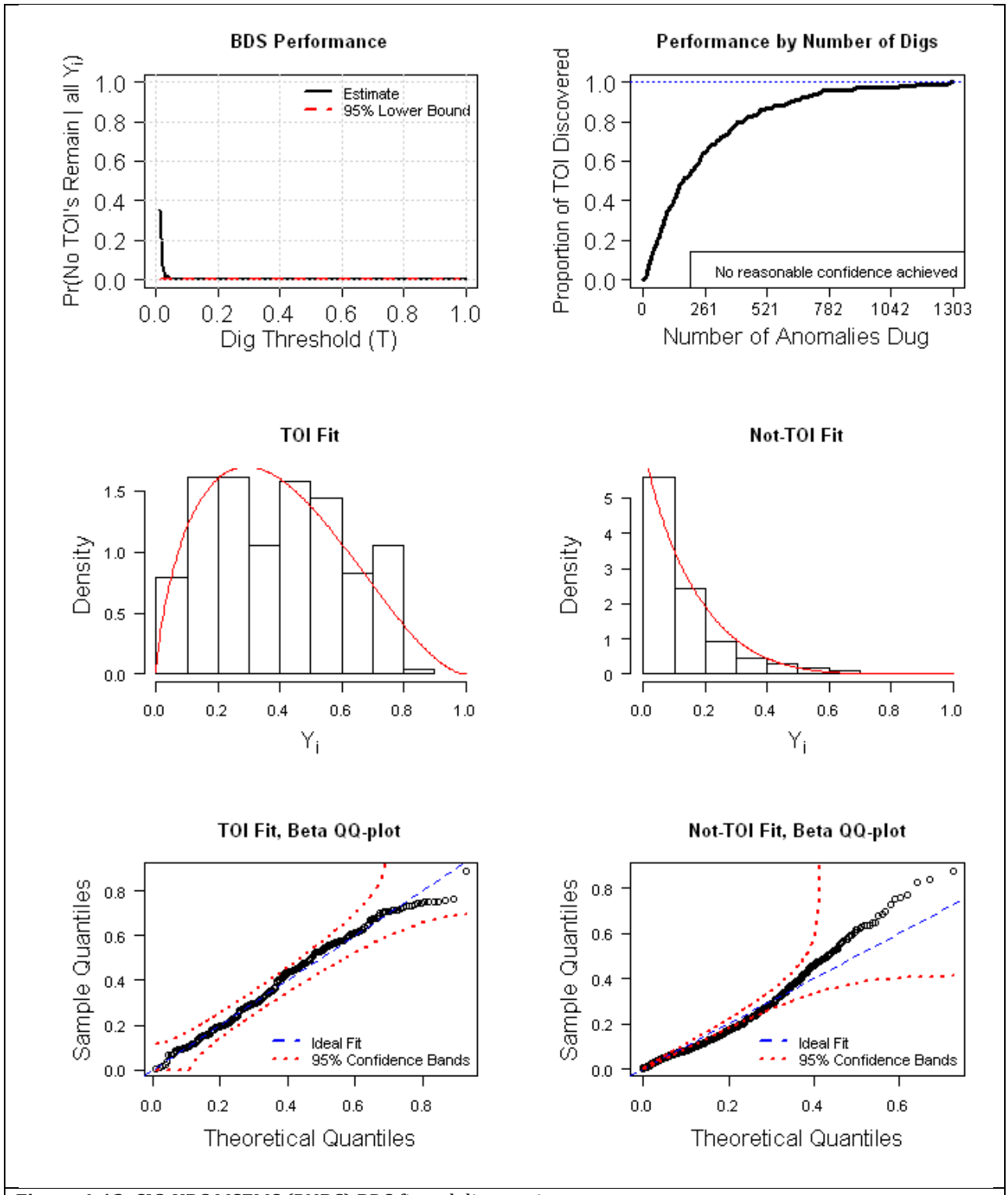

Figure A 12: SIG-UBC MSEMS (PNBC) BDS fit and diagnostics.

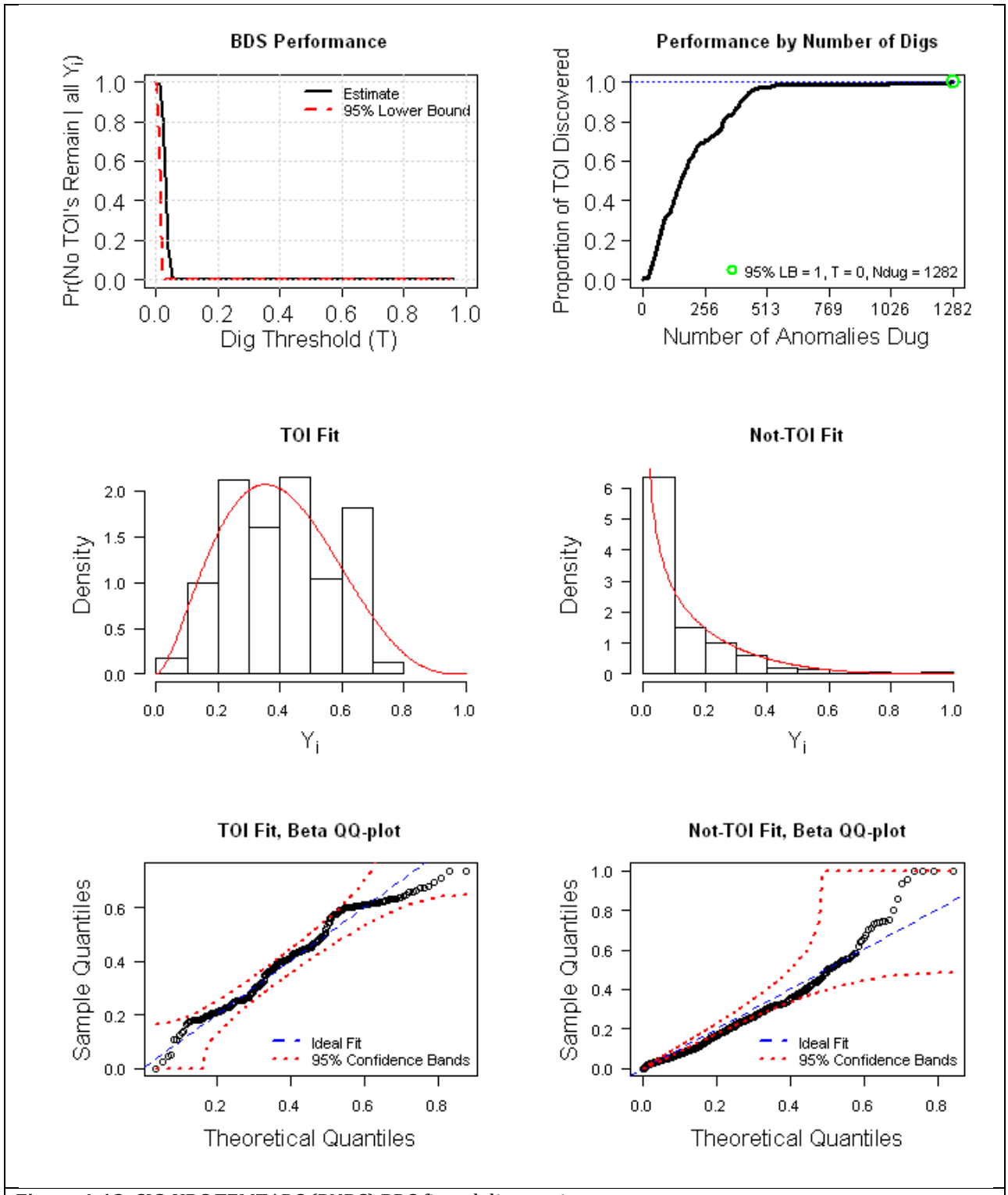

Figure A 13: SIG-UBC TEMTADS (PNBC) BDS fit and diagnostics.

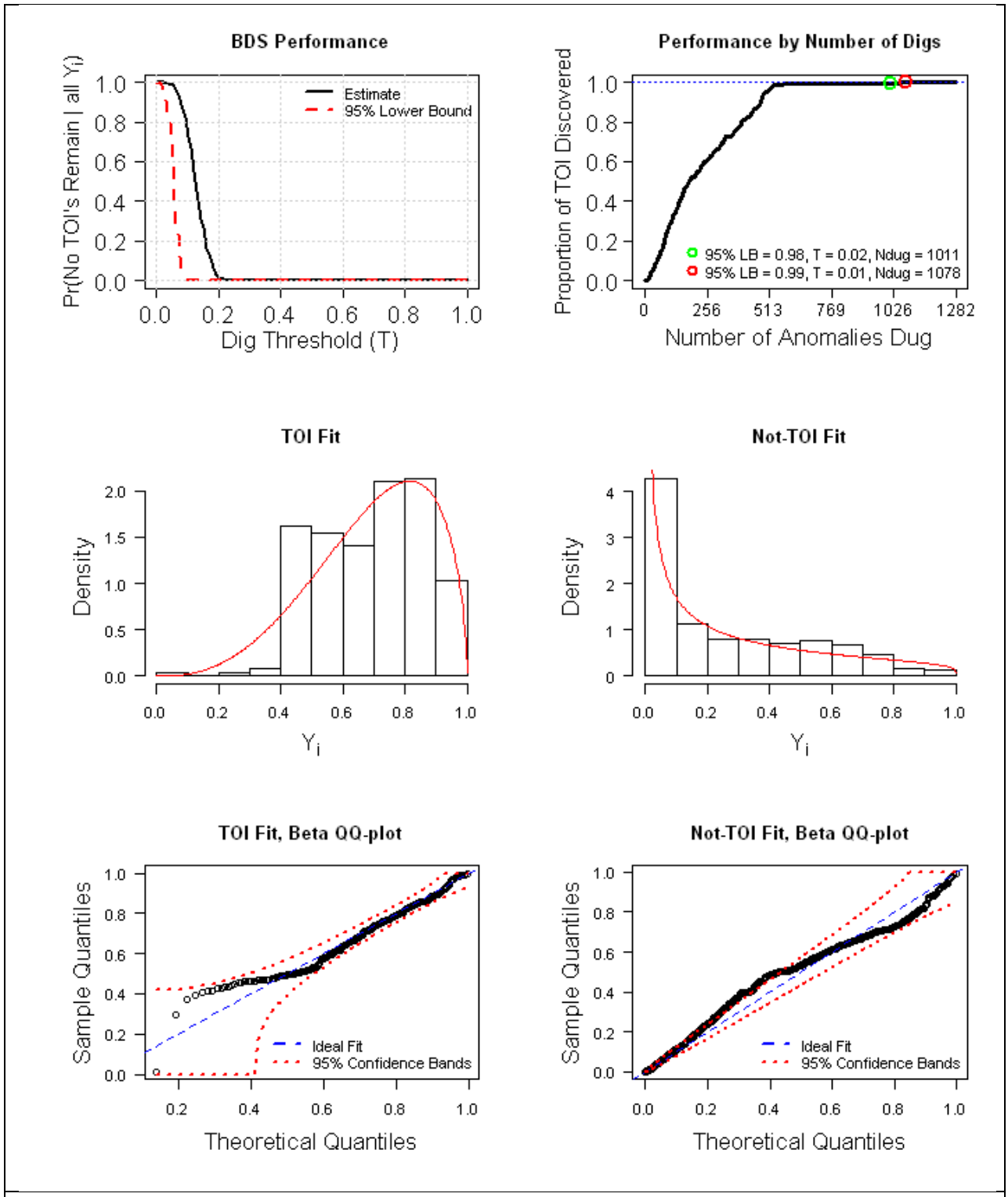

Figure A 14: SAIC UXAnalyze EM61 Array BDS fit and diagnostics.

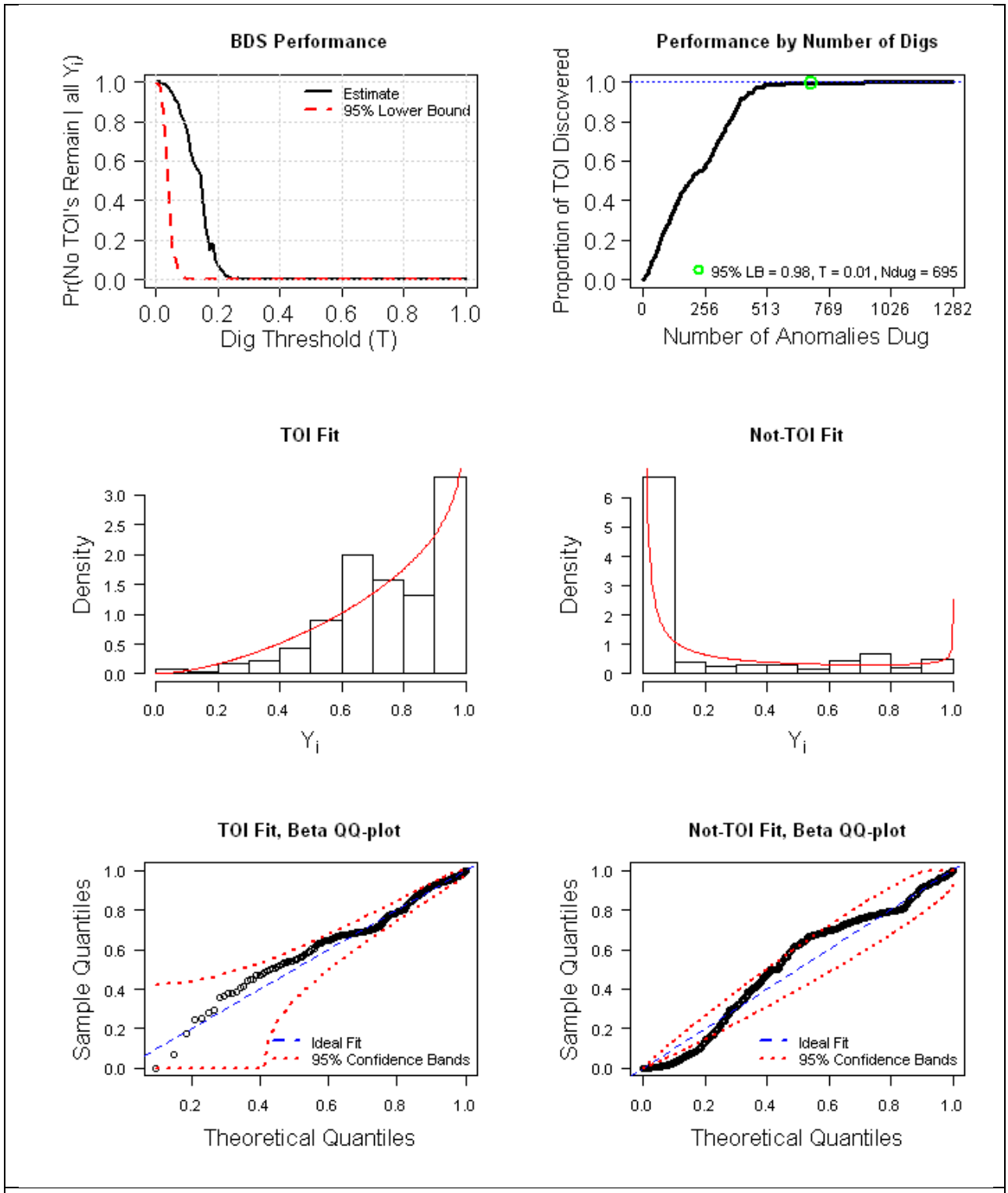

Figure A 15: SAIC EM61 Cart (no slope correction) BDS fit and diagnostics.

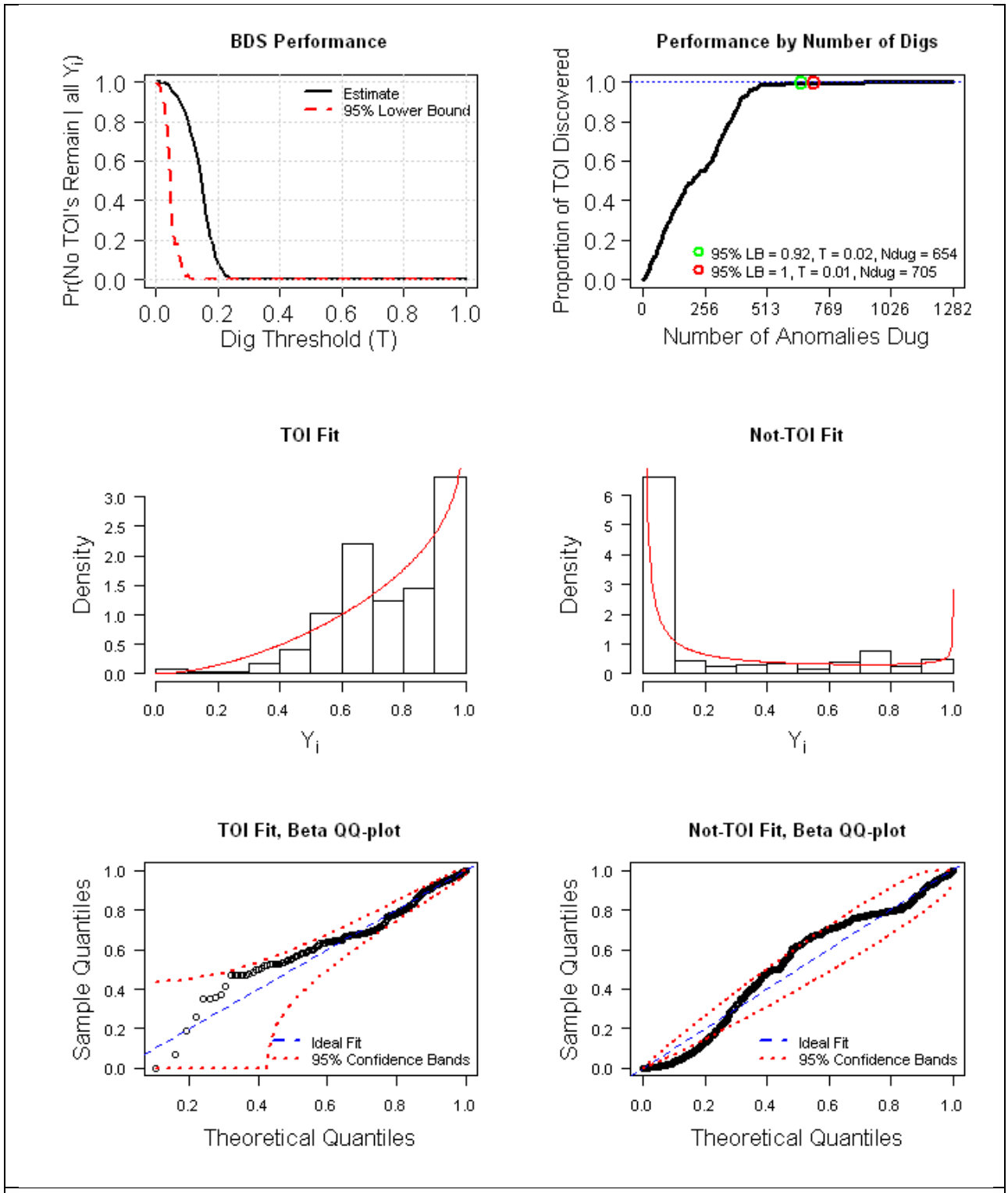

Figure A 16: SAIC EM61 Cart (slope corrected) BDS fit and diagnostics.

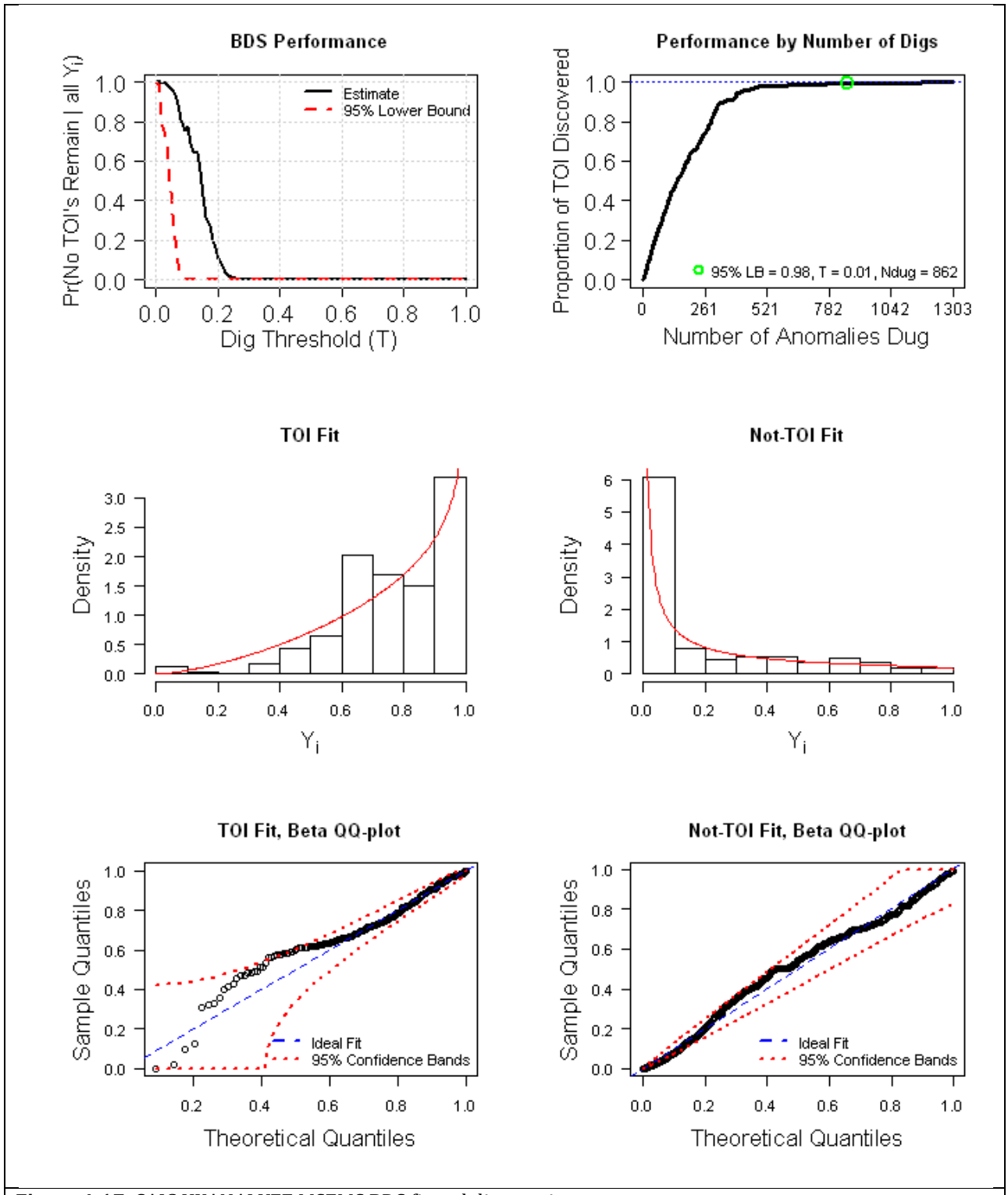

Figure A 17: SAIC UXANALYZE MSEMS BDS fit and diagnostics.

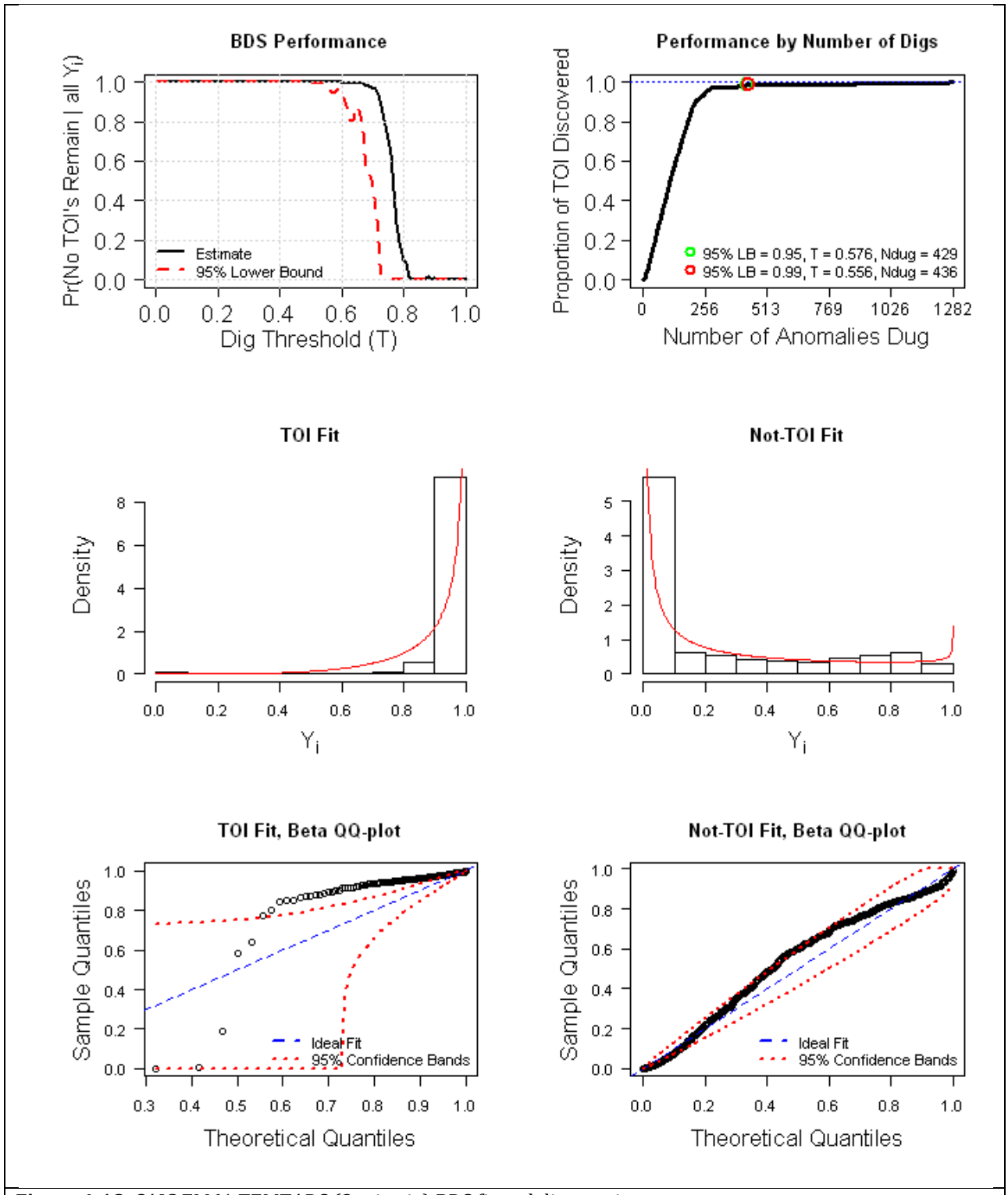

Figure A 18: SAIC EM61 TEMTADS (2 criteria) BDS fit and diagnostics.

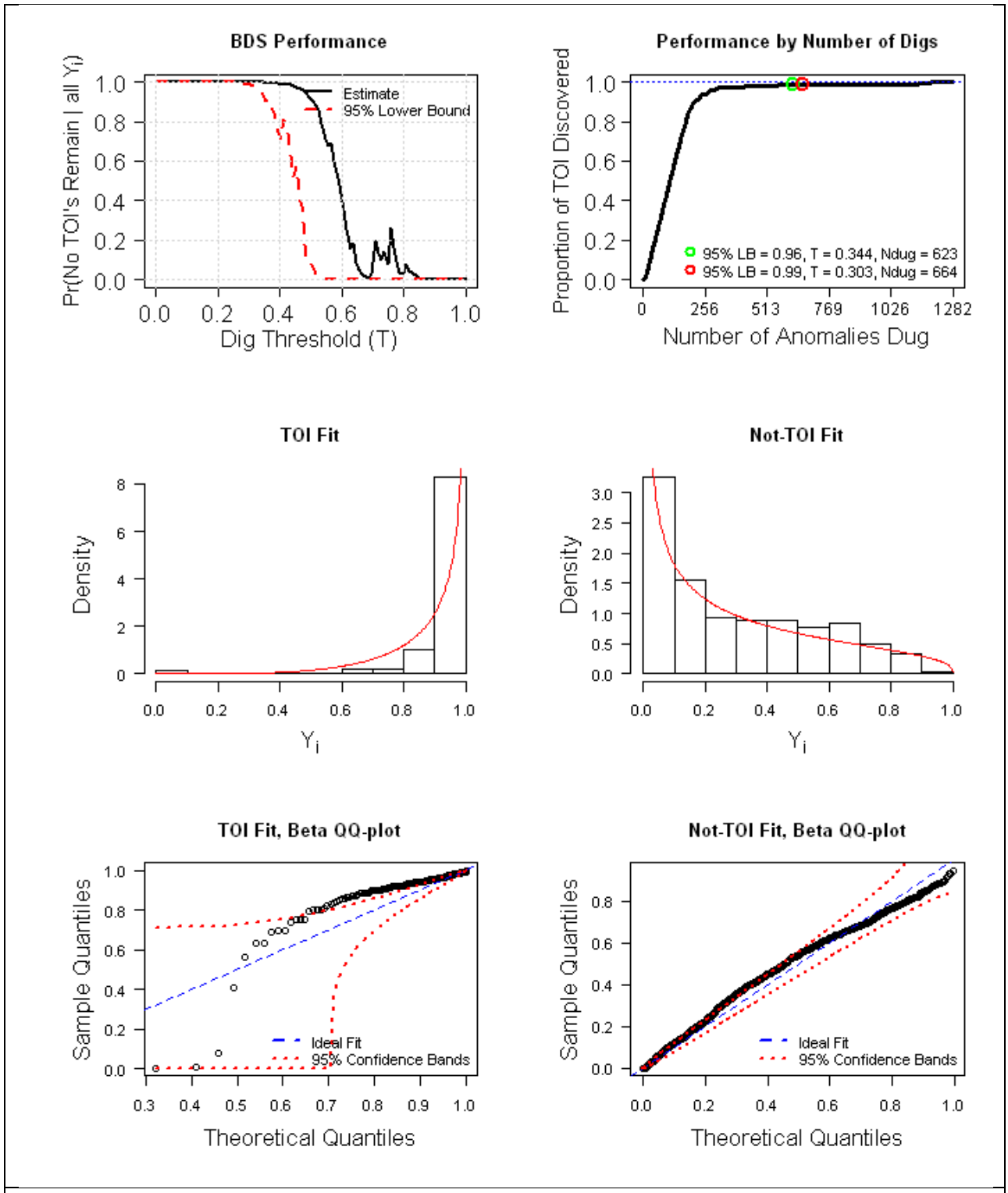

Figure A 19: SAIC EM61 TEMTADS (3 criteria) BDS fit and diagnostics.

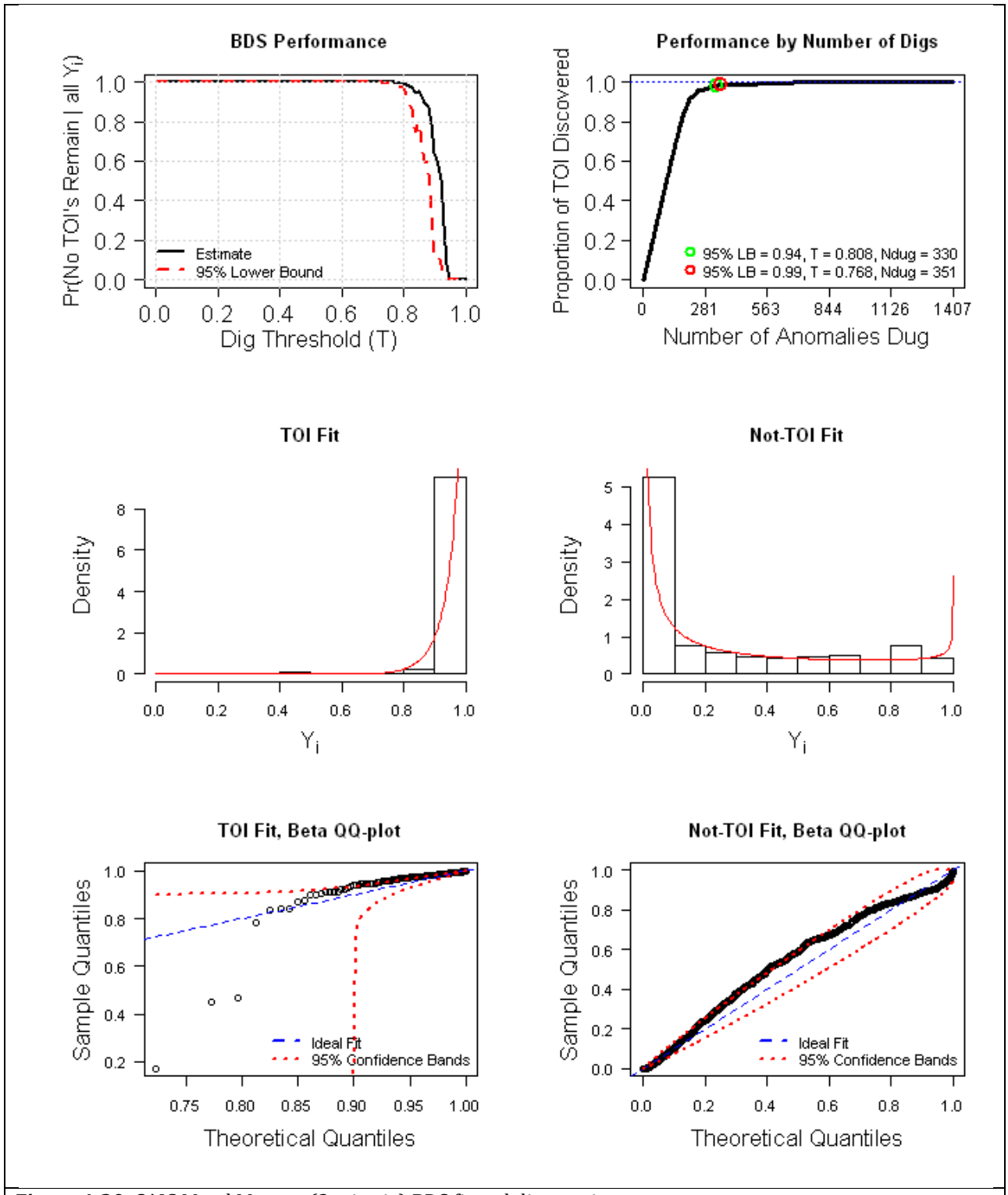

Figure A 20: SAIC Metal Mapper (2 criteria) BDS fit and diagnostics.

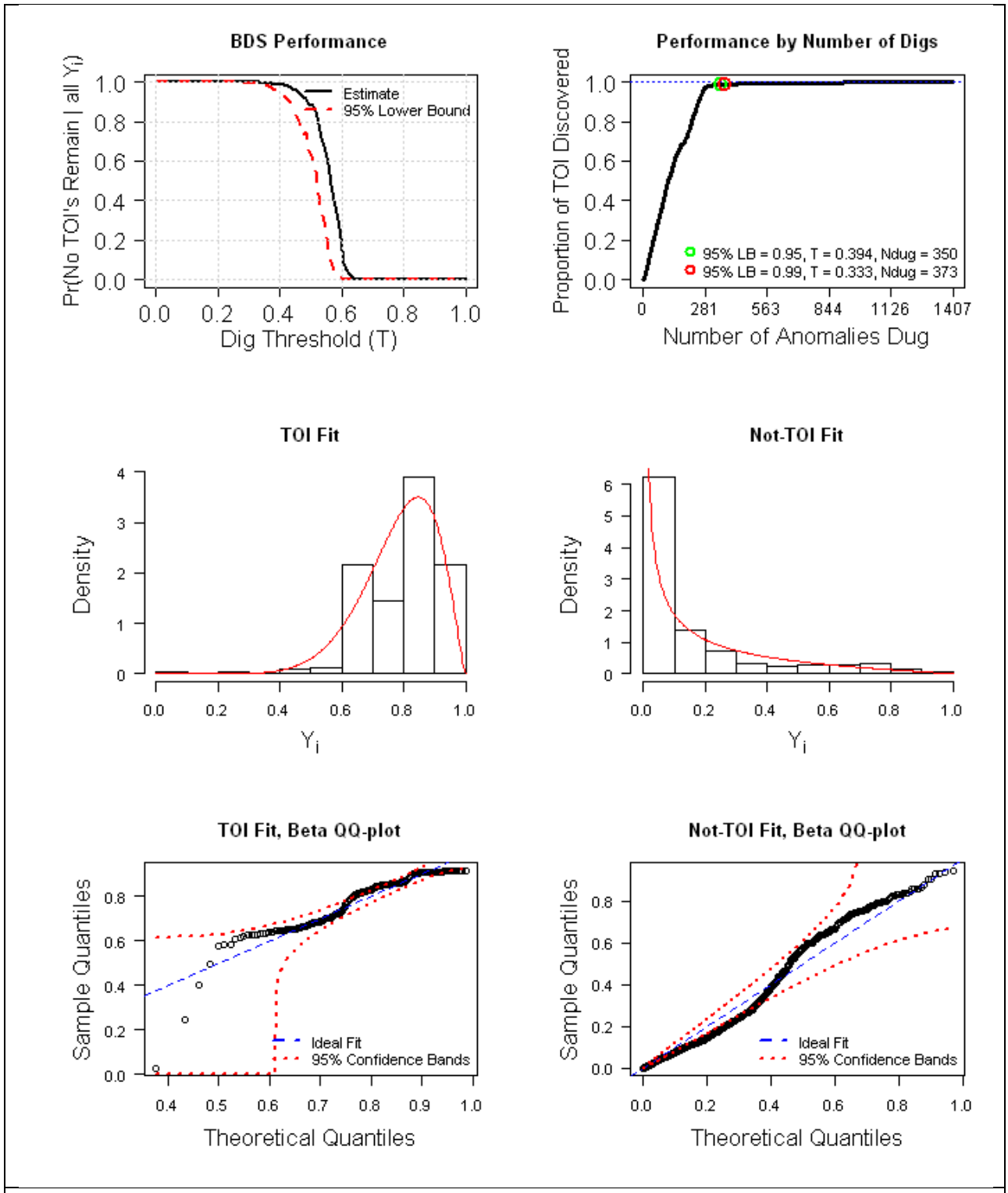

Figure A 21: SKY Metal Mapper BDS fit and diagnostics.

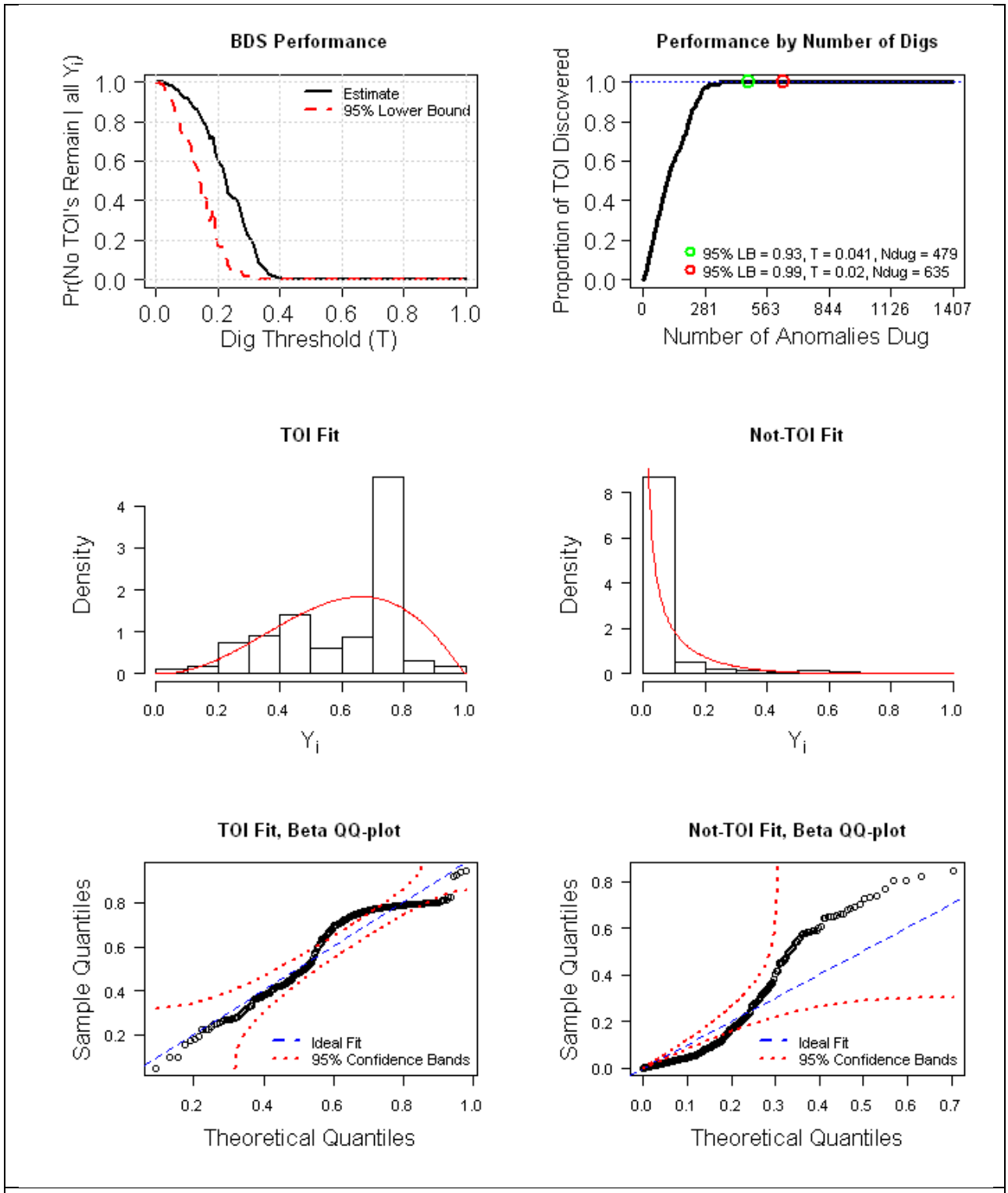

Figure A 22: SIG Metal Mapper (PNBC) BDS fit and diagnostics.

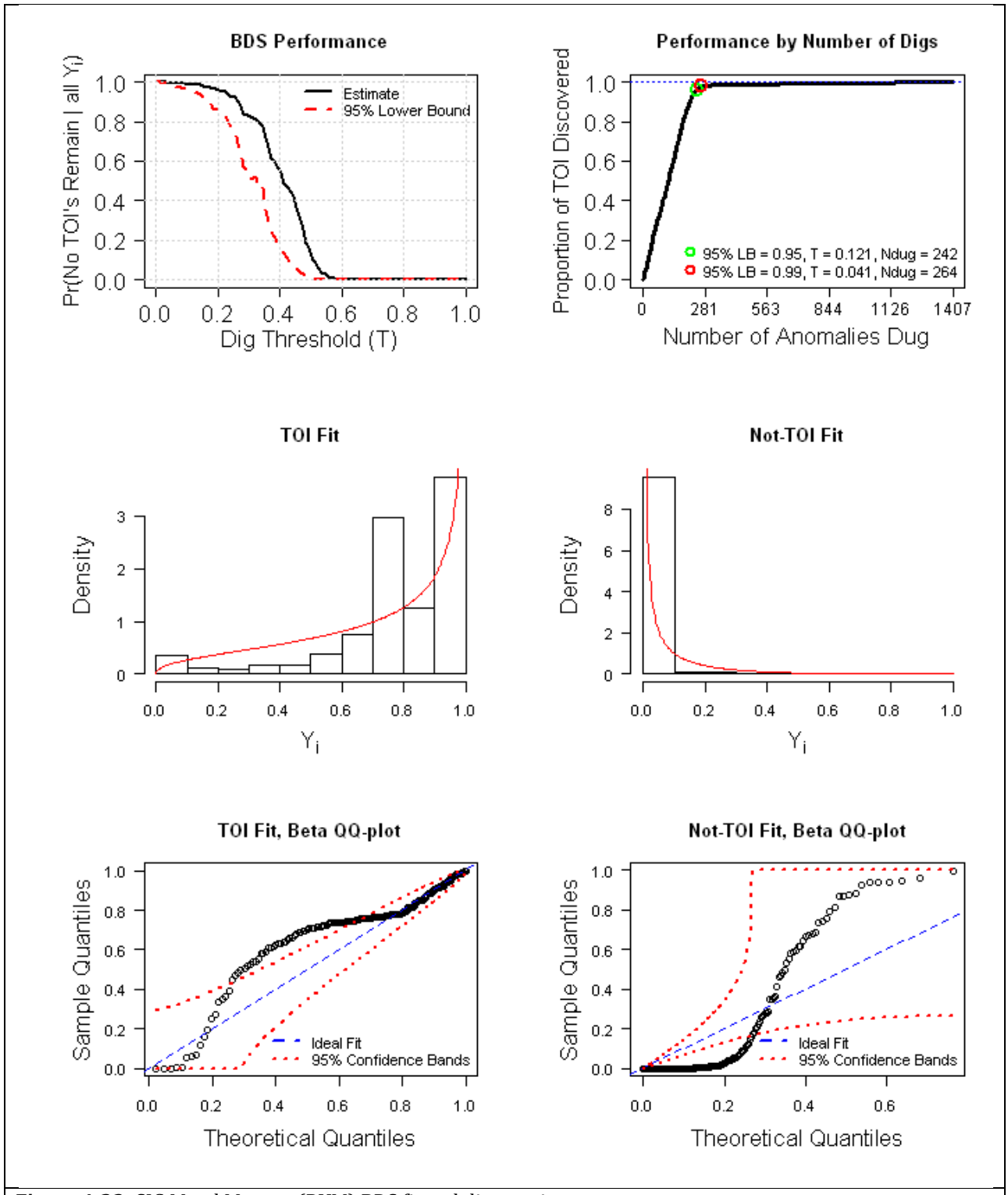

Figure A 23: SIG Metal Mapper (RVM) BDS fit and diagnostics.