Statistical Analysis of SEEM Disaggregation and Reaggregation

Follow Up to BellSouth Statistical Team's Report Filed April 21, 2003

Over the last year an in-depth analysis of the statistical components of BellSouth's Louisiana Self-Effectuating Enforcement Mechanism (SEEM) system has been undertaken jointly by BellSouth Telecommunications, Inc. (BellSouth) and Competitive Local Exchanges Carriers (CLECs). BellSouth filed a report of the analysis on April 21, 2003. That report suggested that more analysis needed to be completed before recommendations concerning changing or not changing the SEEM system should be made.

This report explains the subsequent analyses that have taken place. This report is organized into four sections. Section I provides background information about the SEEM plan and why the Louisiana Public Service Commission (LPSC) staff requested the analysis. A summary of the results of January 2002 through April 2003 performance measurement data analysis is provided in Section II. An outline of BellSouth's and the CLECs' recommendations for future actions is provided in Section III. Section IV provides descriptions of supporting documents that are attached to this report as appendices. There are four appendices attached to this report.

I. Background

SEEM is a system that performs agreed upon calculations in order to assess when the service provided by BellSouth to CLEC customers is as good as the service BellSouth provides to its own customers. When the system's calculations where retail analog standards apply indicate sufficient evidence supporting a disparity in service quality, additional calculations are performed to determine a penalty amount that BellSouth pays. Some of the calculations performed within the SEEM system are based on statistical hypothesis testing methods, and are referred to as *parity testing* calculations.

The parity testing methods in the SEEM plan try to answer the question "Are CLEC customers receiving service that is (significantly) worse than that received by similar BellSouth customers?" In order to do this, performance measurement data first must be disaggregated to insure that CLEC transactions are compared with similar BellSouth transactions (like-to-like comparisons). The statisticians refer to this as disaggregation to the *cell-level*. The cell-level is generally a very deep disaggregation level, and it is not necessarily the level at which parity judgments should be made.

Statistical reaggregation techniques are used within the SEEM system for many reasons that are associated with sound statistical practices. For example, CLEC sample sizes are sometimes very small for individual cells, and this can lead to "noisy" (imprecise) comparisons at the cell-level. In the reaggregation stage, cell-level measures of evidence about the service relative to parity received by CLEC customers (*modified Z-scores*) are combined to produce a single test statistic for a submeasure (*a truncated Z-score*). Comparison of the truncated Z-score with the balancing critical value produces a single compliance determination for a submeasure.

CLECs have voiced concern to the LPSC that the current reaggregation levels used in the SEEM plan potentially mask discrimination. Reaggregation may combine cells that differ substantially from each other in terms of the quality of service received by CLEC customers *relative to the service received by BellSouth customers.* For example, for a given submeasure, assume that CLEC customers with dispatched orders systematically receive better service than that received by BellSouth customers in the corresponding "like-to-like" cells. On the other hand, assume that CLEC customers with non-dispatched orders systematically receive poorer service than BellSouth's customers in the corresponding cells. In this case, there is not a single correct answer to the question posed above (Are CLEC customers receiving service that is (significantly) worse than that received by similar BellSouth customers?). For dispatched orders, the answer is "no," but for non-dispatched orders the answer is "yes."

In response to the CLECs' concerns, the LPSC staff asked a team of statisticians, representing both BellSouth and the CLECs, to review SEEM performance measurement data and determine if there was any statistical evidence of masking that would call for changes in the way the data are disaggregated at the cell-level, and reaggregated at the submeasure level in the parity testing process. Two forms of masking were defined as follows:

> Masking of Discrimination. There is the potential *masking of discrimination* where BellSouth passes the test when the subgroups are not split out, but BellSouth would have failed one of the tests had the subgroups been split out.
>
> Masking of Parity. There is also the potential *masking of parity* where BellSouth fails the test when the subgroups are not split out, but BellSouth would have passed one or both of the tests had the subgroups been split out.

AT&T statistician Dr. Robert Bell represented the CLECs in this process, and BellSouth had PricewaterhouseCoopers LLP statistical consultant Dr. Edward Mulrow, one of the authors of the Louisiana Statisticians' Report,[1] participated in the analysis. A team of statisticians from Ernst & Young LLP, including Dr. Mary Batcher, Ms. Susan Garille Higgins and Ms. Ru Sun, also participated in the analysis, and provided most of the data processing work. BellSouth also requested that Dr. Fritz Scheuren, another author of the Louisiana Statisticians' Report, join the team partway through analysis. Other representatives from BellSouth, AT&T, as well as a LPSC staff representative also provided input at various stages.

There were no statistical tools available to assess whether or not masking occurred in the SEEM system, so the statisticians applied related concepts and developed two diagnostic tools to assess the situation. The main diagnostic studied by the statisticians was a test for heterogeneity. The statisticians used the following definition of heterogeneity:

[1] "Statistical Techniques For The Analysis And Comparison Of Performance Measurement Data." Submitted to the LPSC, Docket U-22252 Subdocket C. Revised February 28, 2000.

Heterogeneity. *Heterogeneity* is a systematic tendency for relative performance on a submeasure to be better for one subset of transactions (group of cells) than for another subset.

Since all cell-level Z-scores are produced on a standardized scale, distinguishing homogeneity and heterogeneity was difficult but in the end turned out to be doable. The team developed a test statistic, Z_{AB}, which is designed to have a standard normal distribution for an arbitrary split of a homogeneous group of cells. However, when heterogeneity exists Z_{AB} should systematically deviate from zero. A diagnostic graphical tool was also developed to assess when masking was taking place. These two diagnostic tools together allowed the statisticians to see if there was any association between heterogeneity and masking. Section II of Appendix 1 provides more detail on these concepts.

Through the use of the Z_{AB} statistic to determine if heterogeneity was present and diagnostic graphical tools, the statisticians explored SEEM performance measure data from the January 2002 through April 2002 time period. This exploration enabled Dr. Bell to lay out a set of criteria for judging when heterogeneity was systematically present. This set of criteria was then applied to the May 2002 through April 2003 time period. Additionally, the diagnostic graphic tools were used to determine when masking was present during the same time period. The results and conclusions of this analysis are presented below.

II. Results

The work done jointly by the CLEC and BellSouth statisticians began with an exploratory phase where the SEEM performance measurement data for the period January 2002 through April 2002 were examined. These results were reported to the LPSC on April 21, 2003.[2] The two diagnostic analysis tools already mentioned were created during this period, but because so little data had been examined, only four months, there was insufficient information to draw conclusions about individual submeasures. A further test of 12 more months was agreed to with data from May 2002 through April 2003. It is the results from these additional 12 months that will be focused on in this report.

As in the original analysis of January 2002 through April 2002 data, only those situations where cell counts were generally[3] at least 20 for the CLECs and BellSouth were

[2] Over the 4-month period from January 2002 through April 2002, there were 128 combinations of measure, mode, factor, and month that are examined. Descriptions of these combinations can be found in Table 1. In over half (57%) of these combinations, there is no heterogeneity detected and in all but 1 of these cases there is no evidence of potential masking. Of the 55 (43%) cases where heterogeneity is detected, 34 (62%) are cases where there appears to be no evidence of potential masking, 6 (11%) cases of potential masking of parity service, and 15 (27%) cases of potential masking of discriminatory service. Of these 15 cases, 9 (60%) occur for PMIA, Mode 1 for various categories. The other 6 cases are distributed more or less evenly among ACNI, MAD, PT30, and RT30.

[3] The January 2002 through April 2002 analysis included one case of a cell count of less than 20: PMIA, Mode 1, Non-Residence, March 2002 had a cell count of 15. The May 2002 through April 2003 analysis

examined.[4] (See Appendix 2.) Table 1 details the cases that meet this criterion. Also, as was done in the analysis of the January 2002 through April 2002 data, since the transaction count in Tier I cells can frequently be less than 20, only Tier II cells were considered for analysis.

Table 1. Measure† – Mode‡ – Factor Combinations Analyzed

Measure	Dispatch Status: Dispatched & Non-Dispatched	Order Type: Change & New or Transfer	Product Group: Residence & Non-Residence
PMIA	Modes 1, 4	Modes 1, 4	Mode 1
ACNI	Modes 1, 4	Modes 1, 4	Mode 1
OCI	Modes 1, 4	Modes 1, 4	Mode 1
PT30	Modes 1, 4	Modes 1, 4	Mode 1
MRA	Modes 1, 4		Mode 1
MAD	Modes 1, 3, 4		Mode 1
RT30	Modes 1, 3, 4		Mode 1
CTRR			Mode 1

† Measure Abbreviations: PMIA = Percent Missed Installation Appointments; ACNI = Average Completion Notice Interval; OCI = Order Completion Interval; PT30 = Provisioning Troubles Within 30 Days; MRA = Missed Repair Appointments; MAD = Maintenance Average Duration; RT30 = Repeat Troubles Within 30 Days; CTRR = Customer Trouble Report Rate.

‡ Mode Abbreviations: Mode 1 = Resale POTS; Mode 2 = Resale Design; Mode 3 = UNE Loops; Mode 4 = UNE Loops and Port Combos; Mode 5 = Interconnection Trunks; Mode 6 = UNE xDSL; Mode 7 = UNE Line Sharing.

To move from the early exploratory phase in our first analysis, Dr. Bell developed a set of criteria to be used to confirm whether heterogeneity existed for a given measure – mode – factor combination. The criteria, which require statistically significant patterns of Z_{AB} values in the anticipated direction, were designed to sharply limit the likelihood of finding heterogeneity where none existed. This confirmatory analysis used May 2002 through April 2003 data to test pre-specified hypotheses suggested by the evidence of heterogeneity in the January 2002 through April 2002 data. The analysis determined that heterogeneity was present for 15 combinations of measure, mode, and heterogeneity factor (e.g., dispatched versus non-dispatched), involving 12 distinct submeasures.[5] (See Appendix 3.)

---

included six cases of cell counts less than 20: PMIA, Mode 1, Non-Residence, December 2002, February 2003, March 2003, and April 2003 had cell counts of 17, 12, 13, and 16, respectively. RT30, Mode 3, Non-Dispatched, May 2002 and December 2002 had cell counts of 16 and 18, respectively.

[4] While analyzing this system over the past few years, Dr. Mulrow determined through computer experiments (that is, statistical simulations) that, for many situations, 20 is an acceptable number of cells in order to have a truncated Z-score without severe skewness problems.

[5] The 15 combinations of measure, mode, and factor (shown in parentheses) are ACNI, Mode 1 (Dispatch Status and Order Type); ACNI, Mode 4 (Dispatch Status and Order Type); PMIA, Mode 1 (Dispatch Status and Order Type); MAD, Mode 1 (Product Group); MAD, Mode 4 (Dispatch Status); MRA, Mode 1 (Dispatch Status); MRA, Mode 4 (Dispatch Status); OCI, Mode 1 (Order Type); PT30, Mode 1 (Order Type); PT30, Mode 4 (Order Type); CTRR, Mode 1 (Product Group); and RT30, Mode 1 (Product Group).

Over the 12-month period from May 2002 through April 2003, there are 384 combinations of measure, mode, factor, and month that are examined. Descriptions of these combinations can be found in Table 1. In over half (65%) of these combinations, there is no heterogeneity detected. In all but one of these cases there is no evidence of potential masking.

Of the 134 (35%) cases where heterogeneity is detected, there are 112 (84%) with no evidence of potential masking, 2 (1%) cases of potential masking of parity service, and 20 (15%) cases of potential masking of discriminatory service. Of these 20 cases, 11 (55%) occur for PMIA, Mode 1 for various categories. The other 9 cases are distributed more or less evenly among ACNI, MAD, MRA, and RT30. In short, of the 384 combinations of measure, mode, factor, and month, there were 21 (5%) cases of potential masking.

For Tier II, a penalty payment is computed only if BellSouth fails for three consecutive months for a given measure – mode combination. The team looked for instances where masking of discrimination eliminated a situation where penalty payments should have been calculated. In other words, were there any combinations of failure and potential masking of discrimination that occurred for three consecutive months? During the 12-month period from May 2002 through April 2003, potential masking of discrimination occurred just once for three consecutive months (November 2002 – January 2003). This was for PMIA, Mode 1, Product Group.

In addition, repeated potential masking of discrimination occurred, although not in two consecutive months, for three of the 12 submeasures identified as heterogeneous. (See Appendix 4.)

- PMIA, Mode 1, Dispatch Status: 3 out of 12 months
- MRA, Mode 4, Dispatch Status: 4 out of 12 months
- RT30, Mode 1, Product Group: 2 out of 12 months.

III Recommendations

The recommendations provided in this section are of two types: (1) Recommendations for changes in the basis system itself, and (2) Recommendations for further research on SEEM.

Action Recommendations

The statisticians agree on the findings reported in the Results section. However, there is a lack of consensus about the appropriate action to recommend based on these results. The table below details areas of agreement and disagreement of various recommendations.

Recommendation #1: Split Three of the Existing Submeasures Further[6]	
Dr. Bell's Position	**Dr. Scheuren's Position**
1. For each of these three submeasures, the current aggregation has masked strong evidence of subparity performance on multiple occasions from May 2002 through April 2003. Depending on the submeasure, truncated Z-score values of less than -2.2 for a pre-specified subgroup of cells were masked two, three, or four times in twelve months. For each submeasure, at least one truncated Z-score value reached -3.30 (corresponding to a P-value of less than 1 in 2,000). Consequently, there is no need for and nothing to gain by continued analysis of more months of data for these submeasures.	1. Dr. Bell's concerns about two and maybe all three of these submeasures may be warranted. This is true despite the fact that the link anticipated between heterogeneity and masking of parity turned out to be weaker than expected. Also, there seems to be little evidence that masking of discrimination for these three measure-mode-factor combinations might become more frequent in the future. In fact, the masking of discrimination for these three became relatively less frequent in the May 2002 through April 2003 period (25%) than it had been in the January 2002 through April 2002 period (42%).
2. Whether it is a good idea to collapse some submeasures is a question that requires business expertise beyond that of the statisticians. Presumably, the decision to create separate submeasures for each of the seven modes was based on a business judgment that these distinct sets of products involved distinct service processes that should not be combined for performance measurement.	2. We agree with Dr. Bell's observation that the decision to create separate new measures, whether by combining them or further splitting them, should be based mainly on a business decision. Therefore, we asked BellSouth to use their business judgment to propose three cases where collapsing three current submeasures would make sense, so as to balance the three measures that might have to be split.[7]
3. On the other hand, data analysis can shed light on the assertion that Recommendation #1 would inappropriately increase the probability of Type I errors (suggesting a need to counter this with the collapse of three pairs of submeasures). Past data indicate that the probability of a Type I error for any of these submeasures has been essentially zero because the truncated Z-score statistic was being driven by a group of cells with very good service (see Appendix 2 and Table 1 of Appendix 4). As long as this remains the case, the only type of error that is possible for the other group of cells is a Type II error. Consequently, there is no need to compensate for any submeasures that are split.	3. We believe BellSouth should prepare to implement Dr. Bell's proposal but worry about the possible increase in Type I error and that is why we are recommending a period in which a compensating change be made to keep the number of measures unchanged. We do not agree with the reasoning underlying Dr. Bell's position regarding Type I error. Instead we feel that a period of further testing, where an alternative is considered alongside what is now being done would be prudent. The key phrase in Dr. Bell's observations is the qualifier to his opinion that begins "as long as this remains the case." Without the presence of further evidence, due diligence would suggest the need to compensate for any measures that are split.

[6] Split three submeasures into six submeasures: (1) Split PMIA–Mode 1 into PMIA–Mode 1–Dispatched and PMIA–Mode 1–Non-Dispatched or into PMIA–Mode 1–New or Transfer Orders and PMIA–Mode 1–Change Orders. (2) Split MRA–Mode 4 into MRA–Mode 4–Dispatched and MRA–Mode 4–Non-Dispatched. (3) Split RT30–Mode 1 into RT30–Mode 1–Residence Products and RT30–Mode 1–Non-Residence Products.

[7] BellSouth proposed that they would (1) collapse the two submeasures Resale POTS (Mode 1) and Resale Design (Mode 2) into Resale and (2) collapse the three submeasures UNE Loops (Mode 3), UNE xDSL (Mode 6), and UNE Line Sharing (Mode 7) into UNE Loops.

Recommendation #2: Split Seven of the Existing Submeasures Further[8]	
Dr. Bell's Position	**Dr. Scheuren's Position**
1. While masking by the formal definition did not occur from May 2002 through April 2003 for any of these submeasures, there were instances of large negative truncated Z-scores in the hypothesized direction that were masked (-2.50 for OCI, Mode 1; -4.22 for PT30, Mode 4; and -3.61 for CTRR, Mode 1). Furthermore, there is the potential for masking in the future. If service deteriorates in coming months, there would be little or no chance to detect it using the current submeasure aggregations. Simply monitoring these submeasures for nine more months means that poor performance could easily go unremedied for a year or more.	1. For these measures masking arguably happened so infrequently that the problem is "in the noise" and not warranting adjustment, unless a wholesale series of changes were to be made. (Potential masking of discrimination did not occur during May 2002 through April 2003 for MAD–Mode 4, OCI–Mode 1, PT30–Mode 1, PT30–Mode 4, CTRR–Mode 1, or MAD–Mode 1. Potential masking of discrimination occurred one time out of 12 months for MRA–Mode 1, Product Group.) We agree that masking may occur in the future but propose that only further regular monitoring be done and that this be done in a timely manner, perhaps quarterly.
2. There is no reason not to split these seven measures. As with the three submeasures listed in Recommendation #1, the risk in terms of increased Type I error is very small. In contrast, two other submeasures for which systematic heterogeneity was observed, ACNI Modes 1 and 4, are excluded from this recommendation because splitting them would increase the probability of Type I errors.	2. There seems to be little evidence that masking of discrimination for these seven measure-mode-factor combinations might become more frequent in the future. In fact, the masking did not occur for these seven in the May 2002 through April 2003 period (0%) as it did in the January 2002 through April 2002 period (4%). To reiterate, only regular monitoring is proposed, using the same approach that was taken on data from January 2002 through April 2003.

[8] (1) Split MAD–Mode 4 into MAD–Mode 4–Dispatched and MAD–Mode 4–Non-Dispatched. (2) Split MRA–Mode 1 into MRA–Mode 1–Dispatched and MRA–Mode 1–Non-Dispatched. (3) Split OCI–Mode 1 into OCI–Mode 1–New or Transfer Orders & OCI–Mode 1–Change Orders. (4) Split PT30–Mode 1 into PT30–Mode 1–New or Transfer Orders & PT30–Mode 1–Change Orders. (5) Split PT30–Mode 4 into PT30–Mode 4–New or Transfer Orders & PT30–Mode 4–Change Orders. (6) Split CTRR–Mode 1 into CTRR–Mode 1–Residence Products & CTRR–Mode 1–Non-Residence Products. (7) Split MAD–Mode 1 into MAD–Mode 1–Residence Products & MAD–Mode 1–Non-Residence Products. Note that this recommendation only discusses masking of discrimination. As noted in the Results Section of this report, masking of parity also occurs for certain submeasures. No recommendations for masking of parity are being proposed.

Research Recommendations

The joint statistical work has been a success and should continue at a modest level, if only as a matter of due diligence. After all, the methods currently used by BellSouth in SEEM to compare the service it provides to its customers with the service that it provides to the CLECs' customers are very complex. Occasionally SEEM appears to have small failures favoring either BellSouth or all CLECs in total. A way to discover and assess these is needed and to determine what (if any) repairs are warranted. To this end there are three specific consensus recommendations offered:

Regular Monitoring. Because of the complexity of SEEM a joint team of BellSouth and CLEC statisticians should monitor results regularly. Every twelve months appears sufficient. Initially this would be done by continuing the current examination of heterogeneity and masking but eventually, depending on the two further recommendations made below the monitoring might shift to other system factors. Short reports from this monitoring would be produced regularly for the LPSC by the joint statistical team.

Tier I Masking. Heterogeneity and masking have only been examined on a subset of Tier II data because there is not a sufficient amount of data at the Tier I level to perform the analysis. But we should be careful in drawing conclusions about Tier I based on Tier II analysis; it does not necessarily follow that heterogeneity and masking exist at the Tier I level even if it exists at the Tier II level. There is a consensus among the statisticians that further work here might be useful, if only to develop new diagnostic tools similar to those employed at the Tier II level in the current analysis.

Distributional Concerns. There are several distributional issues that exist in the current system. For example the current SEEM model assumes a normal distribution of the truncated Z-scores. In fact, the distribution may be skewed. As has been proposed in the past, this should be researched to determine whether this weakness is big enough to warrant a fix. There are many cases where small numbers of cells are employed in the calculations, challenging distributional assumptions. These distributional concerns may need research attention. A systematic research effort on extreme values seems needed. The definition of these anomalies and some root cause analysis should be performed. For example, there are unexpected extreme Z_{AB} values that are frequently observed with the ACNI measure. Each of these examples individually and collectively raises concerns of normality of the test statistic (Z_{AB}) under the null hypothesis and under the alternative hypothesis.

IV. Supporting Documents

The following appendices are supplied as supporting documents to this report:

Appendix 1: Statistical Analysis of SEEM Disaggregation and Reaggregation (*Appendix 1 – LA Stat Analysis Summary-21Apr2003 final - changes accepted.doc*)

This document was filed with the LPSC on April 21, 2003 as the Statistician's Report. It summarizes the results of analysis performed on January 2002 through April 2002 data and was submitted in response to LPSC Docket Number U-22252-C. This document includes the following appendices:

- Appendix A: Louisiana Disaggregation Analysis (*Appendix A-LA Disaggregation-2Apr2003.doc*)
- Appendix B: An Analysis of the Time of Month Characteristic: A Report of Some Work in Progress (*Appendix B-Time of Month Results-4Apr2003.doc*)
- Appendix C: Heterogeneity and Masking Appendix (*Appendix C-Heterogeneity and Masking-15Apr2003.doc)*

Appendix 2: Heterogeneity and Masking May 2002 – April 2003 (*Appendix 2 – Heterogeneity and Masking-2003_1119-DRAFT.doc*)

This document provides details of the analysis of May 2002 through April 2003 data. It is an updated version of the Appendix C-Heterogeneity and Masking-15Apr2003.doc file that was submitted as Appendix C to the April 21, 2003 filing.

Appendix 3: Results of Heterogeneity Assessment Associated with Pre-Specified Hypotheses for May 2002 to April 2003 (*Appendix 3 - Results of Heterogeneity Assessment-2003_0902.doc*)

This document was prepared by Dr. Robert Bell. It summarizes Dr. Bell's assessment of heterogeneity for pre-specified submeasures based on the information provided in Appendix 2: Heterogeneity and Masking May 2002 – April 2003.

Appendix 4: Assessment of Masking for Submeasures Previously Determined to be Heterogeneous (*Appendix 4 - Assessment of Masking-2003_0918.doc*)

This document was prepared by Dr. Robert Bell. It summarizes Dr. Bell's analysis of masking based on the information provided in Appendix 2: Heterogeneity and Masking May 2002 – April 2003.