Simulations of Solid Lumber Strength Property Monitoring Tests

Steve Verrill, Mathematical Statistician
David E. Kretschmann, Research General Engineer
James W. Evans, Mathematical Statistician
Forest Products Laboratory, Madison, Wisconsin

1 Introduction

Dimension lumber, visually graded in accordance with the National Grading Rule and assigned design values derived in accordance with procedures found in ASTM D 1990, has provided satisfactory performance in homes and other structural applications for many years. This ongoing satisfactory performance depends upon a recognition that all standards are living documents that must change over time as new information becomes available. In fact, since its original approval in 1991, D 1990 has been revised to reflect knowledge gained and the changing needs of the wood products industry. These revisions include refinements to the Grade Quality Index provisions, adoption of guiding principles for the maintenance of design values, a framework for monitoring design values, and adoption of a method for deciding when a reassessment of design values is needed. In 2014, a revision to D 1990 was adopted that focused on standardizing the minimum requirements for monitoring, evaluating, and reassessing lumber properties developed in accordance with D 1990. The current paper describes the simulation methods that were used to identify a statistical test that performs acceptably (as determined by the ASTM consensus ballot process) when applied to test data from a monitoring program to determine whether a revaluation of the current resource is needed.

2 Background

It has long been recognized that there is a potential for timber resources to change with time. In the original version of D 1990 (adopted in 1991), Section 13 instructed users of the standard to conduct a reassessment of values derived by the practice if there were cause to believe that there had been a significant change in the raw material resource or product mix. However, no guidance was given on how to monitor or evaluate properties to determine if a reassessment of the allowable properties developed with D 1990 was necessary. In 1994, a methodology for monitoring dimension lumber was proposed (Kretschmann and others, 1999). The methodology was subsequently implemented by the Southern Pine Inspection Bureau. Other agencies implemented their own monitoring efforts. The American Lumber Standard Committee (ALSC, the organization responsible for design value determination in the U.S.) recognized that there was a need for a uniform monitoring approach. Consequently, at the October 12th, 2010 American Lumber Standard Committee Board of Review (ALSC BOR) meeting in Washington DC, a Lumber Property Task Group under the ALSC BOR was reactivated to consider principles for monitoring and evaluating the significance of potential resource changes, and to develop proposed language for changes to ASTM D 1990 that would further explain the process of reassessment described in Section 13 of that practice. This task group was chaired by Charlie Jourdain from the Redwood Inspection Service and was composed of representatives from the major American Lumber Standard rules-writing agencies, and technical representatives from NAHB, FPInnovations, and the USDA Forest Products Laboratory.

Over the course of a year, this group met four times, reviewing and revising several drafts of potential changes to Section 13 of D 1990. The LPTG agreed upon a proposed draft revision to D 1990 that covered the principles of monitoring, evaluating, and reassessing allowable property values after they have initially been determined using D 1990. The LPTG recommended that this revision be taken up by ASTM Section D07.02.01 (solid sawn lumber) as a potential ballot item. A reassessment task group was formed in D07.02.01, and the first ballot on revision of D 1990 Section 13 came out in January 2012. Two approaches were considered for determining whether monitored test cell data suggests that a change in the timber resource has occurred. In one approach, the monitored test data collected for a given size-grade cell is compared with the allowable property value that is claimed for the tested size-grade cell. A second approach compares new size-grade cell data with the size-grade cell data that was used to determine the currently claimed design values. A poll was sent to the Section in August 2012 to determine the preferred approach. The results reviewed at the October 2012 meeting in Madison Wisconsin indicated that the second approach was preferred. It was also clear from this meeting that various methods for performing the comparison needed to be investigated. A task group was formed in ASTM to perform this investigation.

In the spring of 2013 and the summer of 2014, in support of the ASTM task group, USDA Forest Products Laboratory scientists conducted computer simulations to evaluate the performances of a variety of proposed monitoring tests intended to detect reductions in the strength properties of lumber populations. Results from these simulations influenced the content of ASTM standard D 1990's Section 14 and Appendix X11. In this paper, for archival purposes, we describe in some detail the mechanics of the simulations, and the simulation results. We will discuss the simulations in the order in which we conducted them.

Our first set of simulations was based on In-Grade samples and on monitoring samples obtained in 2011/2012. These simulations treated the data sets as simple random samples from lumber populations. These simulations will be discussed in Section 3.

The second set of simulations took into account possible mill and lot effects. That is, they did not treat samples as simple random samples from lumber populations, but instead took into account possible correlations among pieces within a lot or a mill. These simulations were again based on In-Grade data and 2011/2012 monitoring data. They will be discussed in Section 4.

The third set of simulations was based on In-Grade data and new monitoring data obtained in 2014. These simulations also evaluated a proposed new test. They will be discussed in Section 5.

3 In-Grade data and 2011/2012 data. Random effects (mill, lot effects) not considered.

The purpose of these simulations was to compare the performances of a variety of proposed monitoring tests. We compared the tests on the basis of "statistical power." In the context of population monitoring, the "power" of a statistical test is the probability that the test will detect a (real) difference between two populations. In general, statistical power will increase for all reasonable tests as sample sizes increase and as the differences between the populations increase.

To help make "statistical power" clearer: If, for example, there is a 10% difference in the means of two populations, and we repeatedly draw samples of size 100 from each of the two populations, and we repeatedly apply Tests 1 and 2 to these pairs of samples, and in 45% of these draws, Test 1 detects a statistically significant difference between the populations, while in 78% of the draws, Test 2 detects a statistically significant difference, then we say that for detecting a 10% difference in means, Test 2 is more powerful than Test 1. For detecting a 10% difference in means, it is "more sensitive" or a "better instrument."

In support of the monitoring program, we wanted to use the In-Grade and 2011/2012 data, and computer simulations to obtain values such as the 45% and the 78% for a variety of tests in order to identify a "best" test for detecting differences between a base population and a new population. We also wanted to investigate the manner in which power increased as the population being compared to the In-Grade population "morphed" over time into the 2011/2012 population. That is, we wanted to know how the probability of an alarm would increase as the difference between initial and monitored populations increased.

3.1 The data — What were the MOR decreases?

We considered MOR data drawn from 2x4, 2x8, and 2x10 Number 2 and Select Structural populations. To quantify changes from the In-Grade data to the monitoring data, we calculated 5 "global ratios":

1. The ratio of the 2011/2012 sample mean MOR to the In-Grade sample mean MOR.
2. The ratio of the 2011/2012 sample 5th percentile of the MOR to the In-Grade sample 5th percentile of the MOR.
3. The estimated b in the regression model

   $$\text{mornew}_i = b \times \text{morold}_i$$

   where the i denotes the ith order statistic (or estimated order statistic) in the matched samples (see below).
4. The average of the $\text{mornew}_i/\text{morold}_i$ ratios (see below).
5. The median of the $\text{mornew}_i/\text{morold}_i$ ratios (see below).

To obtain ratios 3 through 5, we needed "matching" mornew and morold values. If the new and old sample sizes were the same, we would have simply matched the order statistics (that is, we would have matched the smallest new with the smallest old, ..., matched the largest new with the largest old). However, as one can see from Table 1, the new and old sample sizes do not match. Thus, we first interpolated between order statistics in the larger data set to obtain estimated sample quantiles that matched the sample quantiles of the smaller data set.

For example, there were 413 specimens in the In-Grade ("old") 2x4, Number 2 sample and 408 specimens in the 2011 ("new") 2x4, Number 2 monitoring sample. To obtain appropriate ratios, we first sorted both populations. sortold(i) was the ith order statistic in the sample from the old population. sortnew(i) was the ith order statistic in the sample from the new population. To obtain the reductions in the order statistics in the course of the monitoring, we needed to match up order statistics. We interpolated in the sortold data to obtain estimated order statistics that matched the 408 new order statistics. The ith order statistic in this interpolated data set was called oldest(i). The fractional reduction associated with the ith order statistic in the new population, sortnew(i), was then calculated as

$$\text{ratio}(i) = \text{sortnew}(i)/\text{oldest}(i) \tag{1}$$

For the purposes of the simulations, the expected "full" reduction from the (new) baseline 2011/2012 values (over the course of a monitoring period such as that between the original In-Grade sampling and the 2011/2012 sampling) associated with the ith largest member of the baseline population is then given as either

$$\text{diff}(i) = \text{sortnew}(i) - \text{ratio}(i) \times \text{sortnew}(i) \tag{2}$$

or

$$\mathrm{diff}(i) = \mathrm{sortnew}(i) - \mathrm{rat} \times \mathrm{sortnew}(i) \tag{3}$$

where rat could be, for example, any of the five "global ratios" enumerated above.

If the number of In-Grade values were smaller than the number of corresponding values in the new data set (as in the 2x4, SS and 2x10, Number 2 cases — see Table 1), we interpolated in the sortnew data to obtain newest values that could be matched with the sortold values. In this case

$$\mathrm{ratio}(i) = \mathrm{newest}(i)/\mathrm{sortold}(i) \tag{4}$$

and, for the purposes of the simulations,

$$\mathrm{diff}(i) = \mathrm{newest}(i) - \mathrm{ratio}(i) \times \mathrm{newest}(i) \tag{5}$$

or

$$\mathrm{diff}(i) = \mathrm{newest}(i) - \mathrm{rat} \times \mathrm{newest}(i) \tag{6}$$

If, in the simulations, we used equations (3) or (6) to model the MOR reduction (that is, if we used a global ratio value), we would be assuming that the fractional reduction in MOR was constant across the strength distribution. If we used equations (2) or (5) to model the reduction, we would be assuming that the fractional reduction could change over the strength distribution.

In fact, in the simulations we only considered a constant fractional reduction in five cases — see rows 1,2,5,7, and 9 of Table 4. In all of these cases, the global ratio that we used was the median ratio reported in column 6 of Table 2. In all other cases, we permitted the fractional reduction to vary over the strength distribution (to vary with i). The results that supported the use of the one-sided Wilcoxon test did not depend upon whether we used a constant or a varying reduction.

For the six data sets considered, the five global ratio values are reported in Table 2. It is clear that these reductions are of practical importance, especially for the 2x4, No. 2 population. Related plots are attached as Figures 1 – 12. In odd numbered figures, we plot ordered monitoring data versus ordered In-Grade data (sortnew versus oldest in the cases in which the In-Grade samples were larger than the corresponding monitoring samples, and newest versus sortold in the two cases in which the In-Grade samples were smaller than the corresponding monitoring samples). In the even numbered figures, we plot the (new mor)/(old mor) ratio versus approximate quantile. For example, the leftmost data point in Figure 2 has x coordinate $1/(408 + 1)$ and y coordinate sortnew(1)/oldest(1), and the rightmost has x coordinate $408/(408 + 1)$ and y coordinate sortnew(408)/oldest(408).

3.2 How were the MOR decreases incorporated into the simulations?

As noted above, in addition to comparing the powers of various proposed tests, we wanted to investigate how the powers of these tests would increase as monitored strengths decreased. To do this for cases in which the number of In-Grade strength values was greater than the number of 2011/2012 values (corresponding to equations (1), (2), and (3)), we generated reduced strength data of the form

$$\text{newred}(i) = \text{sortnew}(i) - \delta \times \text{diff}(i) \tag{7}$$

where, in our simulations, we set δ to the values .1, .2, .3, .4, .5, or .6. A δ of, for example, .5 corresponds to a reduction equal to one-half the total reduction observed between the In-Grade data and the 2011-2012 data. In Table 3a, we provide the median strength reductions (as fractions) that correspond to various δ values. In Table 3b, we list the approximate MOR reductions as a function of δ for 2x4, No. 2 data and for 2x8, Select Structural data.

For cases in which the number of In-Grade strength values was less than the number of 2011/2012 values (corresponding to equations (4), (5), and (6)), we generated reduced strength data of the form

$$\text{newred}(i) = \text{newest}(i) - \delta \times \text{diff}(i) \tag{8}$$

The results of the simulations are provided in Table 4 where powers are provided as a function of δ.

3.3 Additional simulation details

For the 2x4, Number 2 simulations discussed in this section, to calculate diff(i) we used either equation (2), or equation (3) with rat equal to the median ratio(i). For the 2x8, SS simulation discussed in this section, to calculate diff(i) we used equation (2). For the 2x10, Number 2 simulations discussed in this section, to calculate diff(i) we used equation (5).

For simulations involving the 2x4, Number 2 data, to draw a sample from the (new) baseline population, 408 numbers were randomly drawn with replacement from 1, ..., 408. If the number i was drawn, for example, 3 times, then sortnew(i) appeared 3 times in the sample drawn from the (new) baseline population.

Similarly, to draw a sample from the (new) reduced population, 408 numbers were randomly drawn with replacement from 1, ..., 408. If the number i was drawn, for example, 2 times, then newred(i) appeared 2 times in the sample drawn from the (new) reduced population.

For the 2x8, SS data, the sample sizes were 409.

For the 2x10, Number 2 data, the sample sizes were 412.

3.4 The proposed tests

3.4.1 One-sided Wilcoxon

The Wilcoxon rank-sum test (also called the Wilcoxon-Mann-Whitney test) is a nonparametric test that is discussed in many introductory statistics texts (for example, Bhattacharyya and Johnson (1977), Miller, Freund, and Johnson (1990), Snedecor and Cochran (1989)), and in all nonparametrics texts (for example, Hollander and Wolfe (1999), Lehmann (1975)). It is used to test whether one population tends to have larger values than another population. It is a simple nonparametic alternative to a Student's t test. The test is available in most major statistical packages, including R and SAS.

We simulated the *one-sided* Wilcoxon test which rejects the null hypothesis of no difference between the monitored and base populations if the rank sum associated with the monitored population is too low.

3.4.2 Sequential One-sided Wilcoxon

This sequential test rejects the null hypothesis of no difference between the base population and the monitored population only if two (one-sided) Wilcoxon tests reject the null hypothesis. (So the procedure would be: Draw a monitoring sample and compare this sample with the base sample via a standard one-sided Wilcoxon test. If the one-sided Wilcoxon test is not statistically significant, do not reject the null hypothesis. If this first test is statistically significant, obtain a second monitoring sample and compare it with the base sample via a standard one-sided Wilcoxon test. If the second test is not statistically significant, do not reject the null hypothesis. If it is statistically significant, do reject the null hypothesis.) To perform this test, we find (via simulation) individual "test sizes" (also known as "alpha levels" and most commonly set to 0.05) so that the overall test size will be approximately equal to 0.05. That is, the probability of a statistically significant result in each of the two possible tests will be larger than 0.05, but the probability that the sequential test leads to a rejection of the null hypothesis when it should not (when there is no difference between the base and monitored populations) will be only (approximately) 0.05. On average, this sequential test will involve more work and be more costly than a single one-sided Wilcoxon test, but it will also be more powerful.

3.4.3 Kolmogorov-Smirnov

The Kolmogorov-Smirnov test is another well-known nonparametric test that can be used to test for a difference between two populations.

3.4.4 Ratio of tolerance bounds

This is an ad hoc test that has been proposed by wood scientists. The idea is that when we are obtaining allowable properties, we begin by obtaining a one-sided, lower 75% nonparametric confidence bound on the 5th percentile of a strength distribution (a 75% confidence, 95% content lower tolerance bound for the strength distribution). (The software that we used to identify the order statistic needed to obtain this confidence bound — confi1.f — can be found at `http://www1.fpl.fs.fed.us/os.html`. Similar software that can be run over the Web can be found at `http://www1.fpl.fs.fed.us/nonpar.html`.) Thus, we are interested in whether this tolerance bound has become "significantly lower." So we might reasonably compare the estimated tolerance bound from a monitoring sample with the estimated tolerance bound from the base sample. We will conclude that a "statistically significant" reduction has occurred if the ratio of the new to old tolerance bounds falls below some pre-set "alarm level." In our simulations we had to determine an appropriate alarm level by drawing the "base" and "monitoring" samples from the same population, and then varying the alarm level until the chance of (falsely) concluding that a reduction had occurred was 0.05. That is, we set the alarm level so that we would have a "false positive" in only 1 of every 20 tests.

3.4.5 Ratio of 5th percentiles

This is an ad hoc test that has been proposed by wood scientists. It is very similar to the ratio of tolerance bounds test discussed above. The only difference is that rather than looking at the ratio of the tolerance bounds estimated from two samples taken from two populations, we look at the ratio of the population fifth percentiles estimated from the two samples. Again, we must first do simulations to obtain "alarm levels" that will yield false positive rates approximately equal to 0.05.

3.4.6 New 5th percentile below one-sided lower 95% confidence bound on old 5th percentile

This is an ad hoc test that has been proposed by wood scientists. It “issues an alarm” if the 5th percentile estimated from the monitoring sample falls below a nonparametric one-sided lower 95% confidence bound on the 5th percentile calculated from the base sample. (This confidence bound equals a particular order statistic [that depends on the sample size] of the base sample. confi1.f [see Section 3.4.4] was used to identify the appropriate order statistic.)

3.4.7 New 5th percentile below two-sided confidence bound on old 5th percentile

This is an ad hoc test that has been proposed by wood scientists. It “issues an alarm” if the 5th percentile estimated from the monitoring sample falls below the left endpoint of the two-sided 95% confidence bound on the 5th percentile calculated from the base sample. (This left endpoint equals a particular order statistic [that depends on the sample size] of the base sample. The program that we used to identify the appropriate order statistic — confi.f — can be found at `http://www1.fpl.fs.fed.us/os.html`.)

3.4.8 χ^2 test based on a combined data, one-sided, lower, 75% confidence bound on the 5th percentile

To perform this test, we do the following: We have n observations from the base sample and n observations from the monitoring sample. We combine these samples and calculate a one-sided, lower, 75% confidence bound, L, on the 5th percentile of the combined population. We then fill a 2x2 table with (in the first row) the number from the base sample that lie below L and the number from the base sample that lie above L, and (in the second row) the number from the monitoring sample that lie below L and the number from the monitoring sample that lie above L. We then perform a chi-squared test on this 2x2 table.

3.5 Discussion of the simulation results

In Table 4, we provide the results from our initial set of simulations.

These simulations involved at least 20,000 trials per test. The uncertainties in the third digit (thousands) of the powers are on the order of .003.

The “med ratio” tests correspond to simulations in which we calculated diff(i) via equation (3), where the rat in the equation was the median of the ratio(i)’s. The “diff for each i” tests correspond to simulations in which we calculated diff(i) via equations (2) or (5).

We only report power values for δ up to 0.6. That is, we only report power results for changes up to a change that amounts to six tenths of the difference observed between the In-Grade population and the 2011/2012 monitoring population. By the time that the change reaches this level, detection by a one-sided Wilcoxon test is almost certain (the Wilcoxon power values exceed 0.99).

The results from the first four lines of the Table correspond to Kolmogorov-Smirnov (K-S) and one-sided Wilcoxon nonparametric tests of the null hypothesis that the baseline and “reduced” populations do not differ. Because these tests were run with a significance level of .05, they would yield a false positive only 5% of the time (on average, once in twenty years if tests are run every year). These simulations suggest that the one-sided Wilcoxon test outperforms the K-S test. That is, when a reduction is present ($\delta > 0$), the one-sided Wilcoxon test generally has a larger probability of detecting it.

In the next four lines we report results from tests in which the hypothesis of no reduction is rejected if the tolerance bound ratio (new/baseline) falls below an alarm level. It is clear that an alarm level of .95 would be unsatisfactory as it would correspond to a 1 in 6 chance of a false alarm. An alarm level of .918 would correspond roughly to a 5% significance level (a 1 in 20 chance of a false alarm). It is clear from the simulations (compare lines 7 and 8 with lines 2 and 4) that, for the same chance of a false positive, the one-sided Wilcoxon test provides better power than does a ratio of tolerance bounds method.

In lines 9 through 12 we consider tests with .025 significance levels. That is, if they are used, there is only a 1 in 40 chance of a false positive. Again, it is clear that the one-sided Wilcoxon test is superior to methods that involve looking at ratios of tolerance bounds or fifth percentiles.

It is clear from lines 13 and 14 that the "estimated new 5th percentile below one-sided lower 95% confidence bound on the old 5th percentile test" (95% confidence, 95% content one-sided, lower tolerance bound test) and the "chi-squared test based on a 75% one-sided lower confidence bound on the 5th percentile estimated from the combined old and new data" test are not competitive with a one-sided Wilcoxon test. Even though the first of these two tests has a size (probability of a false positive) of .1185, it quickly becomes less powerful than a one-sided Wilcoxon test as δ increases. The chi-squared test always has much poorer power than the one-sided Wilcoxon test.

Line 15 of Table 4 applies to 2x8, SS data rather than to 2x4, Number 2 data. Because 2x4, Number 2 MOR values decreased by about 25% over the course of the monitoring period, and 2x8, SS data only decreased by about 10 or 11%, we expected 2x8, SS power to be much lower than 2x4, Number 2 power. It is indeed lower (compare to line 4), but one still obtains very good power (provided that we have a sample size approximately equal to 400) after a strength reduction equal to one-half the total strength reduction between the In-Grade sampling and the 2011/2012 sampling (that is, for $\delta = 0.5$).

Lines 16 and 17 of Table 4 apply to 2x10, Number 2 data. From Figure 10, one can see that the reduction ratio is lower in the left tail of this distribution than elsewhere. Thus, one might believe that in this case, a test based on a ratio of fifth percentiles might be better than a one-sided Wilcoxon test. The simulations reported in lines 16 and 17 of Table 4 do not support this intuition. That is, the power values reported in line 17 for a test based on a ratio of fifth percentiles are considerably lower than the corresponding power values reported in line 16 for a one-sided Wilcoxon test.

Line 18 of Table 4 applies to 2x4, Number 2 data. The line is from the simulation of a sequential test. A second (one-sided Wilcoxon) test is performed if the first (one-sided Wilcoxon) test is statistically significant. Because the results are not statistically independent (because the same "old" data is used in both tests), single tests at a $\sqrt{.05} = .2236$ significance level lead to an overall test at a .103 significance level, not a .05 level. If, however, we run the single tests at a $\sqrt{.0186} = .136$ significance level (found via trial and error), we obtain an overall test with a .0500 (determined via an 80000 trial simulation) significance level. Comparing lines 4 and 18, we see that, as expected, the sequential test yields somewhat better power than a single one-sided Wilcoxon test.

Listings of the computer programs that were used to produce Table 4 can be found at `http://www1.fpl.fs.fed.us/no_mixed_effects.html`

4 In-Grade data and 2011/2012 data. Random effects (mill, lot effects) considered.

In the course of the standards process, Conroy Lum of FPInnovations raised the legitimate point that the simulations described above might not be based on a correct mixed effects model. That is, they did not take into account potential region effects, mill within region effects, and lot within mill effects (as well as piece to piece variation within a lot). In response to this criticism we fit mixed effects models to In-Grade 2x4, Number 2 data and 2011 2x4, Number 2 data.

For the In-Grade 2x4, Number 2 data we found:

1. There is no statistically significant region effect. (It should be noted that analyses performed during Part I of the In-Grade program did detect a statistically significant region effect. Also, in recent

analyses the authors of the current report did detect a statistically significant region effect in 2011 2x8, Number 2 data. The existence of such effects suggests that a two-way nonparametric ANOVA [with the two "treatments" composed of baseline data and monitoring data, and the regions as blocks] might be preferred over a one-sided Wilcoxon test, and, of course, a correct mixed effects analysis would likely yield better results than either a one-sided Wilcoxon test or a nonparametric two-way analysis.)

2. There is no statistically significant lot effect.

3. The mill effects are approximately normally distributed.

4. In general, we cannot statistically reject normality for the (within mill) piece effects.

5. The within mill piece standard deviations are roughly constant across mills.

For the 2011 2x4, Number 2 data we found:

1. There is no statistically significant region effect.

2. It was not possible to estimate lot effects because there was only one lot per mill.

3. The mill effects are approximately normally distributed.

4. In general, we cannot statistically reject normality for the (within mill) piece effects.

5. The within mill piece standard deviations are roughly constant across mills.

Given the mixed effects analyses, we altered our simulations to include random mill effects.
The appropriate statistical model in this case is

$$y_{ij} = \mu + \gamma_i + \epsilon_{ij} \tag{9}$$

where y_{ij} is the MOR of the jth piece in the ith mill, μ is the overall average MOR, γ_i is the random deviation associated with the ith mill, and ϵ_{ij} is the random deviation associated with the jth piece from the ith mill.

In Table 5, for both old and new 2x4, number 2 MOR data, we report the estimates of μ, σ_γ, and σ_ϵ from the linear mixed effects analyses. (σ_γ is the estimated standard deviation of the random mill effect, γ. σ_ϵ is the estimated standard deviation of the random piece effect, ϵ.)

Thus, we have

$$\mu_{\text{new}}/\mu_{\text{old}} = 4.225/5.597 = 0.755 \tag{10}$$

$$\sigma_{\gamma,\text{new}}/\sigma_{\gamma,\text{old}} = 0.5473/0.5944 = 0.921 \tag{11}$$

$$\sigma_{\epsilon,\text{new}}/\sigma_{\epsilon,\text{old}} = 1.729/2.066 = 0.837 \tag{12}$$

The modified simulation proceeded as follows:

For each δ ($\delta \in 1, \ldots, 6$), we performed 80,000 trials. In each trial, we first drew 41 (or 36 or 24 or 12) mill γ_i values from a normal distribution with mean 0 and standard deviation $\sigma_{\gamma,\text{old}}$. For each mill, we drew 10 ϵ_{ij} values from a normal distribution with mean 0 and standard deviation $\sigma_{\epsilon,\text{old}}$. Then the 410 (or 360 or 240 or 120) "old" MOR values in the trial were the

$$\mu_{\text{old}} + \gamma_i + \epsilon_{ij}$$

values.

The "new" values had the form:

$$\mu_{\text{old}} + \delta \times (\mu_{\text{new}} - \mu_{\text{old}}) + \gamma_i + \epsilon_{ij}$$

Thus μ was reduced by the term $\delta \times (\mu_{\text{new}} - \mu_{\text{old}})$. Also, the γ_i's were drawn from a normal distribution with reduced standard deviation

$$\sigma = \sigma_{\gamma,\text{old}} + \delta \times (\sigma_{\gamma,\text{new}} - \sigma_{\gamma,\text{old}})$$

and the ϵ_{ij}'s were drawn from a normal distribution with reduced standard deviation

$$\sigma = \sigma_{\epsilon,\text{old}} + \delta \times (\sigma_{\epsilon,\text{new}} - \sigma_{\epsilon,\text{old}})$$

After the old and new samples for the trial were drawn, a one-sided Wilcoxon test was performed. The power reported in Table 6 is the fraction of 80000 trials in which the null hypothesis of no difference between the old and new populations was rejected.

We can see from Table 6 that, as Dr. Lum suggested, because the ten samples from a particular mill share a γ_i, they are correlated and the one-sided Wilcoxon test is not entirely correct. In particular the α level (or "test size" or "producer's risk") is inflated from the nominal 0.05 to 0.10. (This doubling is coincidental rather than theoretical.) This inflation of producer's risk could be reduced by operating at a smaller nominal α level, or by performing a correct mixed effects analysis.

The authors of the revised ASTM standard chose to adopt a one-sided Wilcoxon approach despite the mixed effects nature of the data. Apparently, they made this choice for two reasons: the relative ease of use of a one-sided Wilcoxon test, and the fact that the standard permits a retest (via a draw of a second monitoring sample). This reduces the overall probability of a false positive.

Listings of the R and SAS codes that were used to perform the mixed effects analyses, and the FORTRAN code that was used to perform the mixed effects simulations can be found at
`http://www1.fpl.fs.fed.us/mixed_effects.html`

5 In-Grade data and 2014 data. Random effects (mill, lot effects) not considered.

In 2014, the authors were asked to review a new 2x4, Number 2 data set and a "new" proposed monitoring test. The test was the "new 5th percentile below two-sided confidence bound on old 5th percentile" test discussed in Section 3.4.7. We place "new" in quotation marks because the proposed test was similar to the one-sided test discussed in Section 3.4.6 and evaluated in the simulations that produced Table 8.

In Table 7, we replicate Table 1 and add the appropriate line for the 2014 data. In Table 8, we replicate Table 2 and add the appropriate line for the 2014 data.

Figures 13 and 14 are the 2014 versions of Figures 1 through 12. That is, in Figure 13 we plot 2014 order statistics versus the corresponding interpolated In-Grade order statistics, and in Figure 14 we plot the ratios of 2014 order statistics to the corresponding interpolated In-Grade order statistics. It is clear from Table 8 and Figures 13 and 14 that the 2014 data behaves differently from the 2011/2012 data. The 2011/2012 data sets display fairly large and relatively constant (within a data set) fractional decreases in strength. But the 2014 data shows a fairly small average decrease in strength, and non-constant decreases across the range of order statistics (fairly large fractional decreases for low and high order statistics, and actual *increases* for intermediate order statistics [data in the middle of the strength distributions]). These changes are reflected in the relative performances of the one-sided Wilcoxon and "5th percentile" tests displayed in Table 9 (which is an extension of a portion of Table 4 — lines 1, 3, and 5 of Table 9 repeat the information in lines 4, 15, and 16 of Table 4). The proposed new test does not perform as well as the one-sided Wilcoxon test (the new test has lower power) in comparing In-Grade data with 2011/2012 data. However, the new test performs better than the one-sided Wilcoxon test in comparing In-Grade data to 2014 data. This would be expected from Figure 14. The reductions (from In-Grade data) in the 2014 data are concentrated at low and high order statistics (in the tails of the strength distributions) and thus it is reasonable that a test that focuses on differences in the lower tail (as does the proposed test) would outperform a global test such as the Wilcoxon.

We note that the 2011/2012 data contained a higher proportion of prime lumber (lumber biased toward the pith) than did the 2014 data. This might help explain the differences between the 2011/2012 data and the 2014 data.

A listing of the computer program that was used to produce Table 9 can be found at `http://www1.fpl.fs.fed.us/no_mixed_effects.html`

6 Summary

We have provided a detailed discussion (and links to source code) of simulations that we performed in support of the ASTM D07.02.01 reassessment task group. These simulations influenced the decision of the task group to recommend the use of the one-sided Wilcoxon statistical test as a lumber strength monitoring tool. This recommendation was subsequently incorporated into ASTM standard D 1990.

In the course of developing and conducting the simulations we found:

1. Comparing Ingrade data with 2011/2012 monitoring data (and ignoring the possibility of region, mill, and lot effects), we found that a one-sided Wilcoxon test performed better than a number of proposed competitors.

2. Performing mixed effects analyses with Ingrade data and 2011/2012 monitoring data, we saw evidence of mill effects but not of lot effects. In some cases we saw evidence of regional effects. Given a "correct" mixed effects simulation, we found that a one-sided Wilcoxon test has an actual significance level that exceeds the nominal significance level. This increases the probability of a false positive when testing for strength reductions.

3. Comparing Ingrade 2x4, 2x8, and 2x10 Number 2 and Select Structural data with 2011/2012 monitoring data, we saw relatively constant fractional reductions in the lower tails, middle, and upper tails of strength distributions. See Figures 2, 4, 6, 8, 10, and 12. In contrast, the decrease in MOR from Ingrade data to 2014 monitoring data (2x4, Number 2) was concentrated in the lower and upper tails of the distribution. See Figure 14. Because of this concentration, the Wilcoxon test did not perform as well as a competing test in detecting differences between the Ingrade data and the 2014 monitoring data.

Our simulation results suggest that a Wilcoxon test is not always optimal. A mixed effects analysis would likely do a better job of protecting against false positives. Also, the analyses reported in Section 5 suggest that if reductions are similar to those seen from the Ingrade data to the 2014 data (concentrated in the tails) rather than to those seen from the Ingrade data to the 2011/2012 data (relatively constant fractional reductions across the whole strength distribution), then a test that was particularly sensitive to changes in the lower tail (and possibly insensitive to changes elsewhere in the distribution) might be preferred. Thus, as additional monitoring experience is gained, it is possible that alternate tests might become more attractive, and revisions to the standard might be justified.

References

ASTM. 2015. Standard practice for establishing allowable properties for visually-graded dimension lumber from in-grade tests of full-size specimens. D 1990-14. West Conshohocken, PA: ASTM International.

Bhattacharyya, G.K. and Johnson, R.A. (1977), *Statistical Concepts and Methods*, New York: John Wiley and Sons.

Hollander, M. and Wolfe, D.A. (1999), *Nonparametric Statistical Methods*, New York: John Wiley and Sons.

Kretschmann, D.E., Evans, J.W., and Brown, L. (1999), *Monitoring of visually graded structural lumber*, Research Paper FPL-RP-576, Madison, WI: U.S. Department of Agriculture, Forest Service, Forest Products Laboratory, 18 pages.

Lehmann, E.L. (1975), *Nonparametrics: Statistical Methods Based on Ranks*, San Francisco: Holden-Day.

Miller, I., Freund, J.E., Johnson, R.A. (1990), *Probability and Statistics for Engineers*, Englewood Cliffs, New Jersey: Prentice Hall.

Snedecor, G.W. and Cochran, W.G. (1989), *Statistical Methods*, Ames: Iowa State University Press.

	Sample size		Matched values	
Size, grade	In-Grade	2011/2012	in plots	
2x4, No. 2	413	408	oldest	sortnew
2x4, SS	413	420	sortold	newest
2x8, No. 2	1367	420	oldest	sortnew
2x8, SS	626	409	oldest	sortnew
2x10, No. 2	412	420	sortold	newest
2x10, SS	413	410	oldest	sortnew

Table 1: Sample sizes

Size, grade	Ratio of means	Ratio of 5th percentiles	b in mornew = $b \times$ morold	average of new/old quantile ratios	median of new/old quantile ratios
2x4, No. 2	.754	.711	.765	.742	.726
2x4, SS	.870	.818	.874	.866	.872
2x8, No. 2	.873	.896	.867	.883	.871
2x8, SS	.890	.883	.888	.892	.895
2x10, No. 2	.833	.741	.836	.826	.838
2x10, SS	.865	.888	.867	.863	.864

Table 2: Five measures of the MOR reduction (the five "global ratios") from In-Grade values to 2011/2012 monitoring values

	δ									
Size, grade	.1	.2	.3	.4	.5	.6	.7	.8	.9	1.0
2x4, No. 2	.973	.945	.918	.890	.863	.836	.808	.781	.753	.726
2x4, SS	.987	.974	.962	.949	.936	.923	.910	.898	.885	.872
2x8, No. 2	.987	.974	.961	.948	.936	.923	.910	.897	.884	.871
2x8, SS	.989	.979	.969	.958	.947	.937	.926	.916	.905	.895
2x10, No. 2	.984	.968	.951	.935	.919	.903	.887	.870	.854	.838
2x10, SS	.986	.973	.959	.946	.932	.918	.905	.891	.878	.864

Table 3a: For the simulations: Median ratios of corresponding order statistics as a function of δ (when using equations (3) or (6))

	δ						
Size, grade	0	.1	.2	.3	.4	.5	.6
2x4, No. 2	0	40	80	120	160	200	240
2x8, SS	0	25	50	75	100	125	150

Table 3b: For the 2x4, No. 2 and 2x8, SS simulations: approximate MOR reduction (in psi) as a function of δ.

Line	Test	δ						
		0	.1	.2	.3	.4	.5	.6
1	red ratio = med ratio Kolmogorov-Smirnov, .05	.0495	.1767	.4819	.8013	.9651	.9969	.9999
2	red ratio = med ratio Wilcoxon, .05	.0497	.2425	.5756	.8678	.9812	.9986	1.0000
3	red ratio = diff for each i Kolmogorov-Smirnov, .05	.0498	.1826	.4963	.8077	.9672	.9979	1.0000
4	red ratio = diff for each i Wilcoxon, .05	.0495	.2280	.5423	.8341	.9671	.9968	.9997
5	red ratio = med ratio ratio of tol bounds, alarm = .95	.1692						
6	red ratio = diff for each i ratio of tol bounds, alarm = .95	.1716						
7	red ratio = med ratio ratio of tol bounds, alarm = .918	.0509	.1554	.2871	.5375	.7154	.8793	.9636
8	red ratio = diff for each i ratio of tol bounds, alarm = .918	.0514	.1590	.2911	.5739	.7497	.8999	.9717
9	red ratio = med ratio Wilcoxon, .025	.0248	.1523	.4546	.7883	.9608	.9962	.9999
10	red ratio = diff for each i Wilcoxon, .025	.0244	.1412	.4185	.7429	.9385	.9925	.9997
11	red ratio = diff for each i ratio of tol bounds, alarm = .897	.0251	.0735	.1925	.3613	.6348	.7930	.9392
12	red ratio = diff for each i ratio of 5th perc, alarm = .908	.0253	.0839	.2424	.4264	.6969	.8643	.9723
13	red ratio = diff for each i est new 5th below 95% tol bound on old 5th?	.1185	.2925	.4689	.7234	.8703	.9812	.9956
14	red ratio = diff for each i χ^2 test based on comb 75% tb on 5th, .05	.0486	.0996	.1956	.4306	.6201	.8952	.9748
15	red ratio = diff for each i 2x8, SS, Wilcoxon, .05	.0496	.1736	.3908	.6653	.8647	.9666	.9941
16	red ratio = diff for each i 2x10, 2, Wilcoxon, .05	.0497	.1862	.4306	.6927	.8901	.9747	.9965
17	red ratio = diff for each i 2x10, 2, ratio of 5th perc alarm = .851, .05	.0508	.0914	.1527	.2339	.3275	.4360	.5595
18	red ratio = diff for each i 2x4, 2, Wilcoxon sequential, overall .0500	.0500	.2513	.6091	.8832	.9834	.9989	.9999

Table 4: Power (probability of rejecting the null hypothesis). The first 14 lines are appropriate for 2x4, Number 2 data. The 15th line is appropriate for 2x8, SS data. The 16th and 17th lines are appropriate for 2x10, Number 2 data. The 18th line corresponds to a sequential one-sided Wilcoxon test.

data	μ	σ_γ	σ_ϵ
old, 2x4, 2, MOR	5.597	0.5944	2.066
new, 2x4, 2, MOR	4.225	0.5473	1.729

Table 5: Estimates of the parameters of the mixed effects models.

	δ						
Sample sizes	0	.1	.2	.3	.4	.5	.6
410	.10	.28	.55	.80	.94	.99	.999
360	.10	.27	.52	.76	.92	.98	.997
240	.10	.23	.42	.64	.82	.93	.978
120	.10	.19	.30	.45	.61	.75	.857

Table 6: For 2x4, number 2 MOR, the power (probability of rejecting the null hypothesis) of a one-sided Wilcoxon test run at a nominal α (producer's risk) level of 0.05.

	Sample Size		Matched Values in Plots	
Size, grade	In-Grade	Monitoring		
2x4, No. 2	413	408	oldest	sortnew
2x4, SS	413	420	sortold	newest
2x8, No. 2	1367	420	oldest	sortnew
2x8, SS	626	409	oldest	sortnew
2x10, No. 2	412	420	sortold	newest
2x10, SS	413	410	oldest	sortnew
2014 2x4, No. 2	413	362	oldest	sortnew

Table 7: Sample sizes

Size, grade	Ratio of means	Ratio of 5th percentiles	b in mornew = $b \times$ morold	average of new/old quantile ratios	median of new/old quantile ratios
2x4, No. 2	.754	.711	.765	.742	.726
2x4, SS	.870	.818	.874	.866	.872
2x8, No. 2	.873	.896	.867	.883	.871
2x8, SS	.890	.883	.888	.892	.895
2x10, No. 2	.833	.741	.836	.826	.838
2x10, SS	.865	.888	.867	.863	.864
2014 2x4, No. 2	.961	.897	.950	.968	.986

Table 8: Five measures of the MOR reduction ratio

		δ							
Line	Test	0	.1	.2	.3	.4	.5	.6	1.0
1	red ratio = diff for each i 2x4, 2, Wilcoxon, .05	.0495	.2280	.5423	.8341	.9671	.9968	.9997	
2	red ratio = diff for each i 2x4, 2, CI, .05	.0571	.1689	.3056	.5774	.7633	.9473	.9812	
3	red ratio = diff for each i 2x8, SS, Wilcoxon, .05	.0496	.1736	.3908	.6653	.8647	.9666	.9941	
4	red ratio = diff for each i 2x8, SS, CI, .05	.0566	.0894	.1122	.1387	.1673	.2262	.2565	
5	red ratio = diff for each i 2x10, 2, Wilcoxon, .05	.0497	.1862	.4306	.6927	.8901	.9747	.9965	
6	red ratio = diff for each i 2x10, 2, CI, .05	.0537	.1157	.1650	.2720	.3734	.4811	.5926	
7	red ratio = diff for each i 2x4, 2, 2014, Wilcoxon, .05	.0539	.0608	.0674	.0769	.0870	.0987	.1109	.1867
8	red ratio = diff for each i 2x4, 2, 2014, CI, .05	.0567	.0921	.1381	.1985	.2933	.3763	.4934	.8698

Table 9: Power (probability of rejecting the null hypothesis). For lines 1,2,7,8, these values are based on 20,000 trials per cell. For lines 3,4,5,6, these values are based on 80,000 trials per cell. (Lines 1, 3, and 5 of Table 9 repeat the information in lines 4, 15, and 16 of Table 4.)

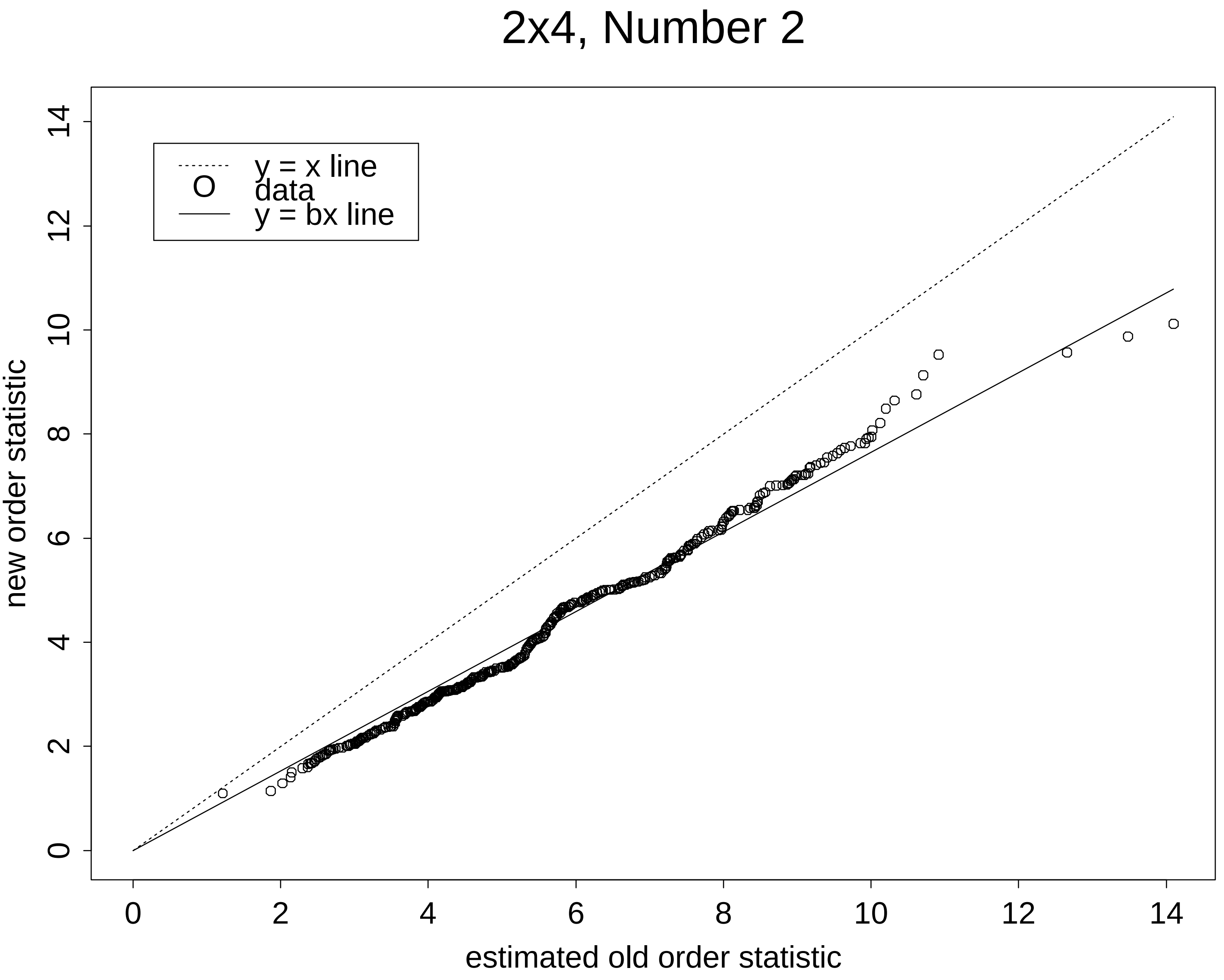

Figure 1: 2x4, Number 2 data. Ordered monitoring data versus estimated ordered original data. The dotted line is the y = x line. The solid line is the fitted sortnew = b*oldest line.

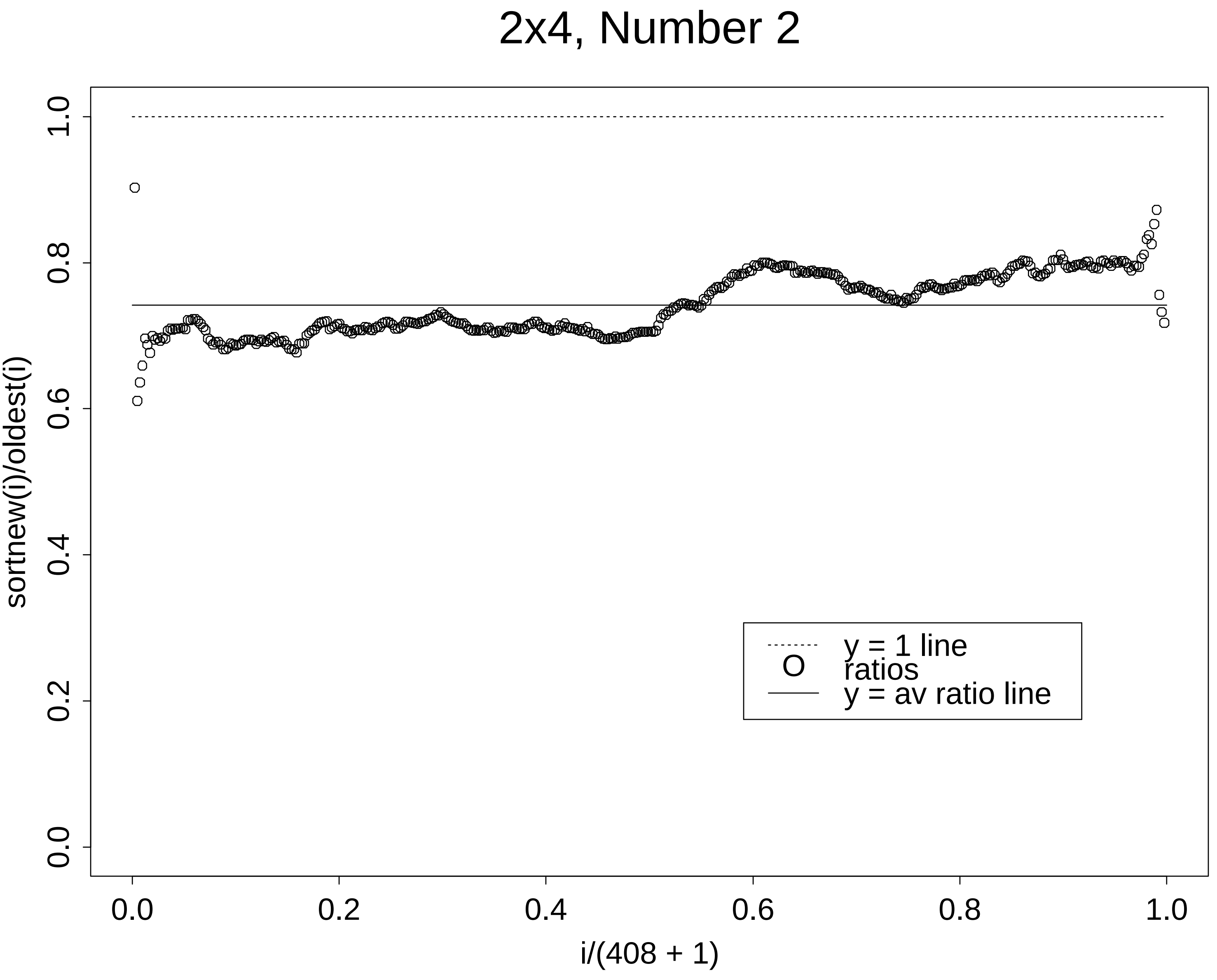

Figure 2: 2x4, Number 2 data. sortnew(i)/oldest(i) versus $i/(408 + 1)$. The dotted line is the y = 1 line. The solid line is the y = average ratio line.

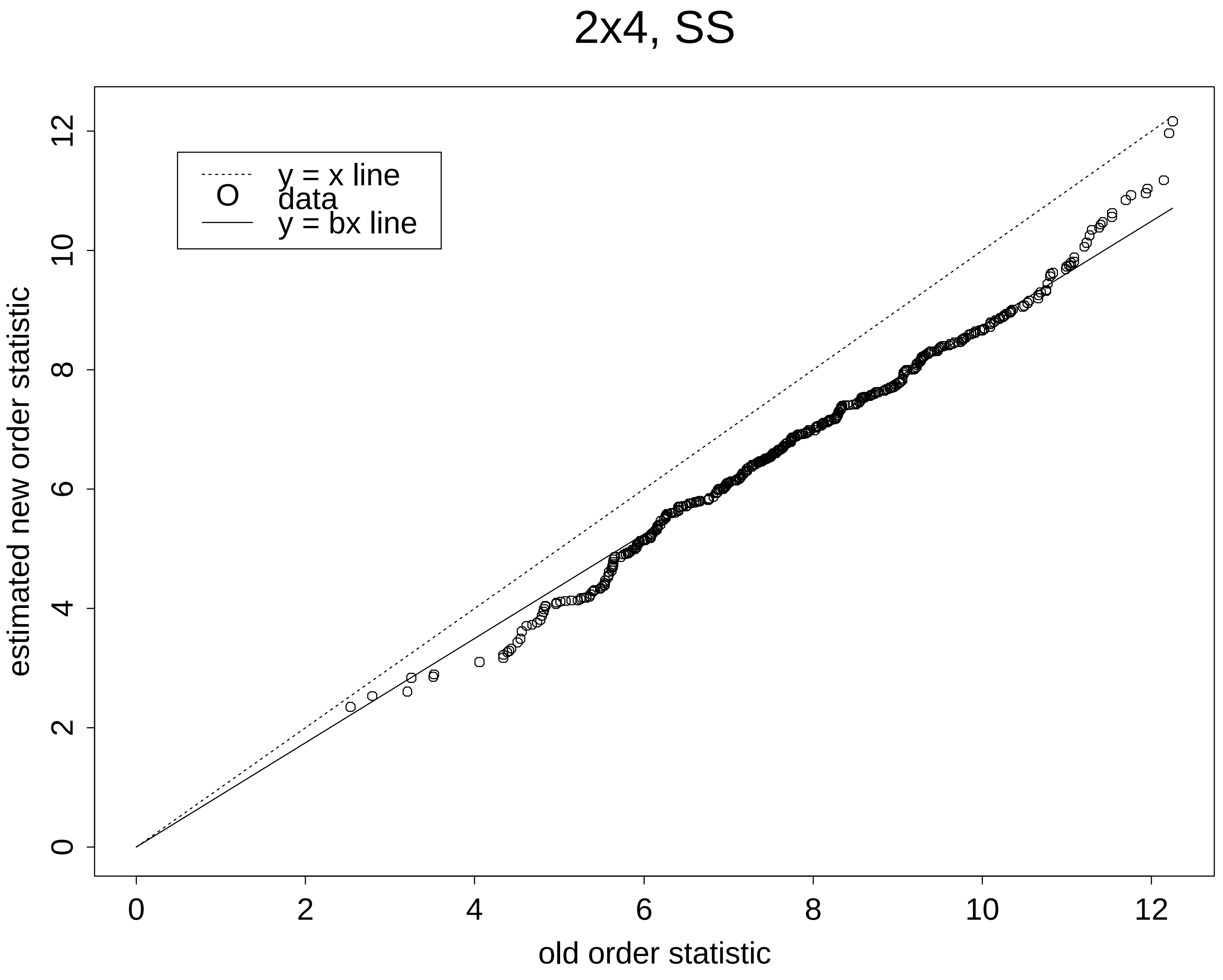

Figure 3: 2x4, SS data. Estimated ordered monitoring data versus ordered original data. The dotted line is the y = x line. The solid line is the fitted newest = b*sortold line.

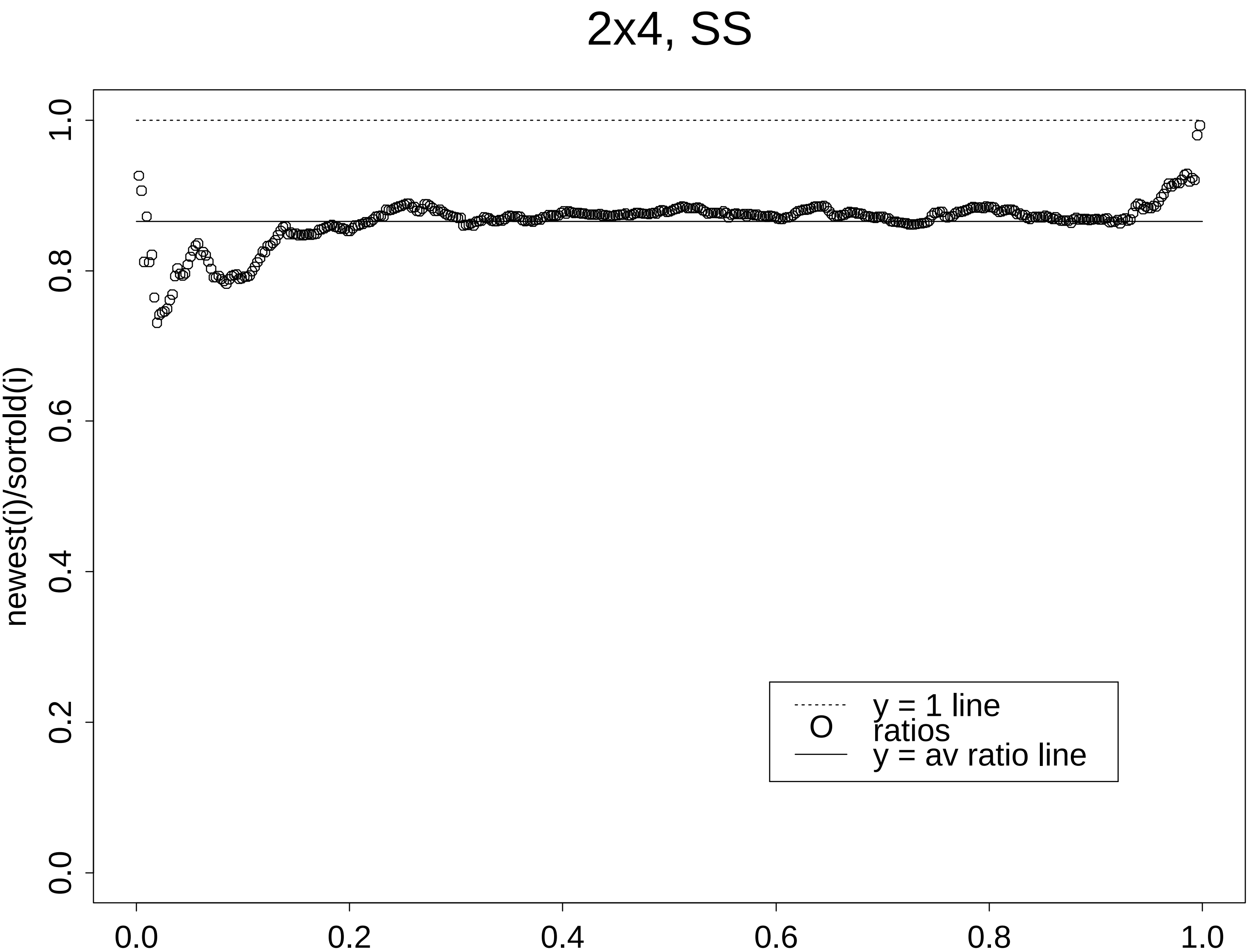

Figure 4: 2x4, SS data. newest(i)/sortold(i) versus $i/(413 + 1)$. The dotted line is the y = 1 line. The solid line is the y = average ratio line.

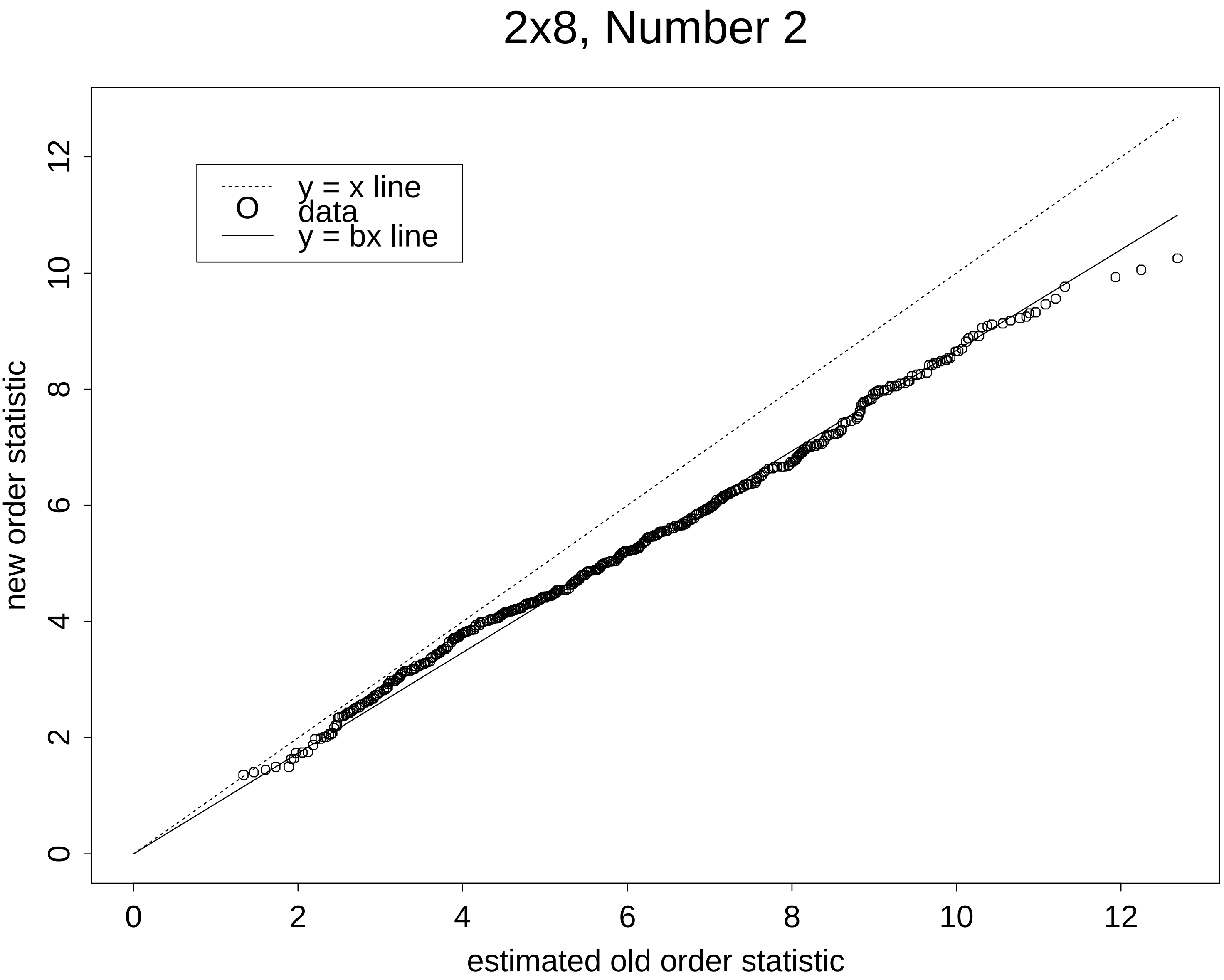

Figure 5: 2x8, Number 2 data. Ordered monitoring data versus estimated ordered original data. The dotted line is the y = x line. The solid line is the fitted sortnew = b*oldest line.

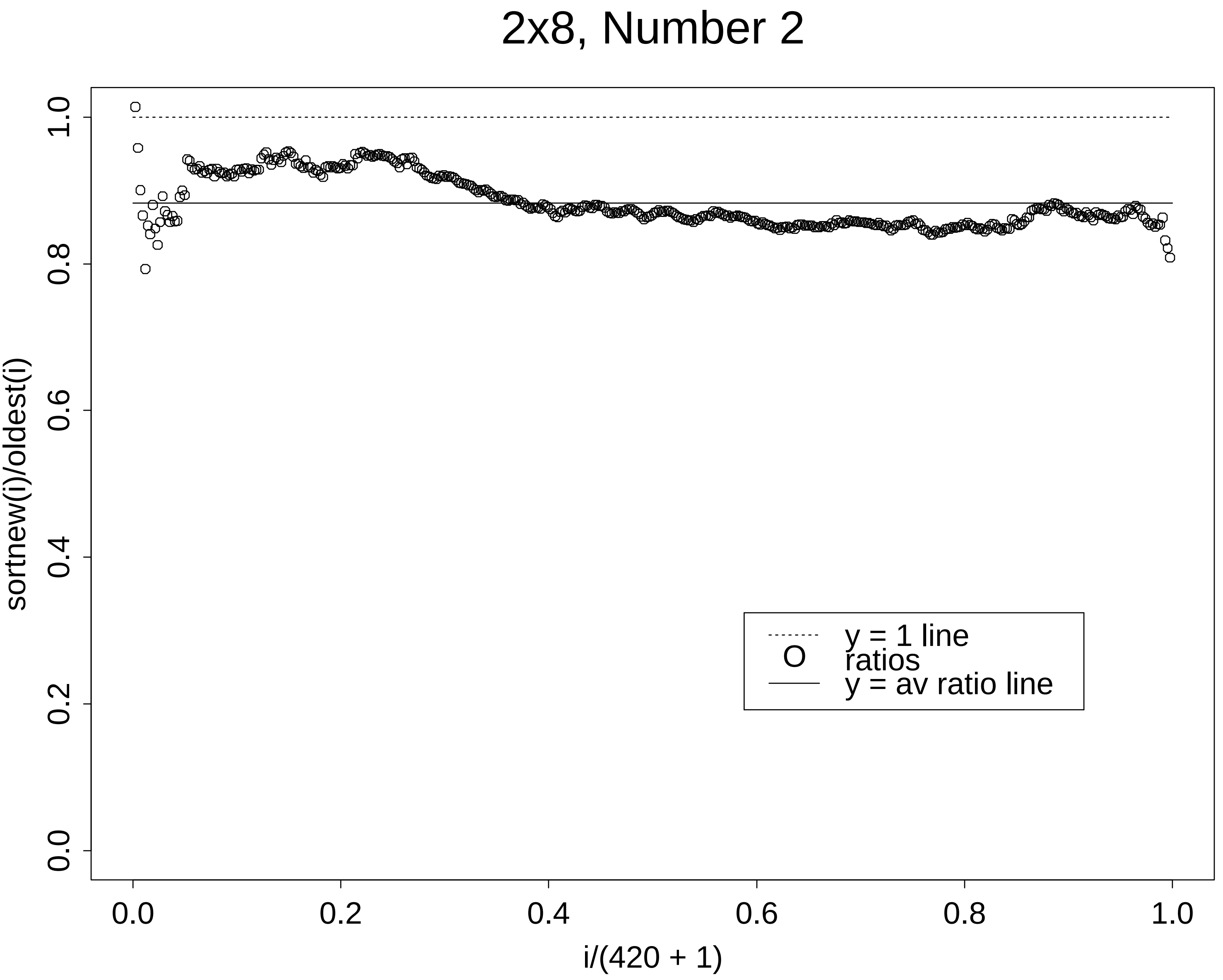

Figure 6: 2x8, Number 2 data. sortnew(i)/oldest(i) versus $i/(420 + 1)$. The dotted line is the y = 1 line. The solid line is the y = average ratio line.

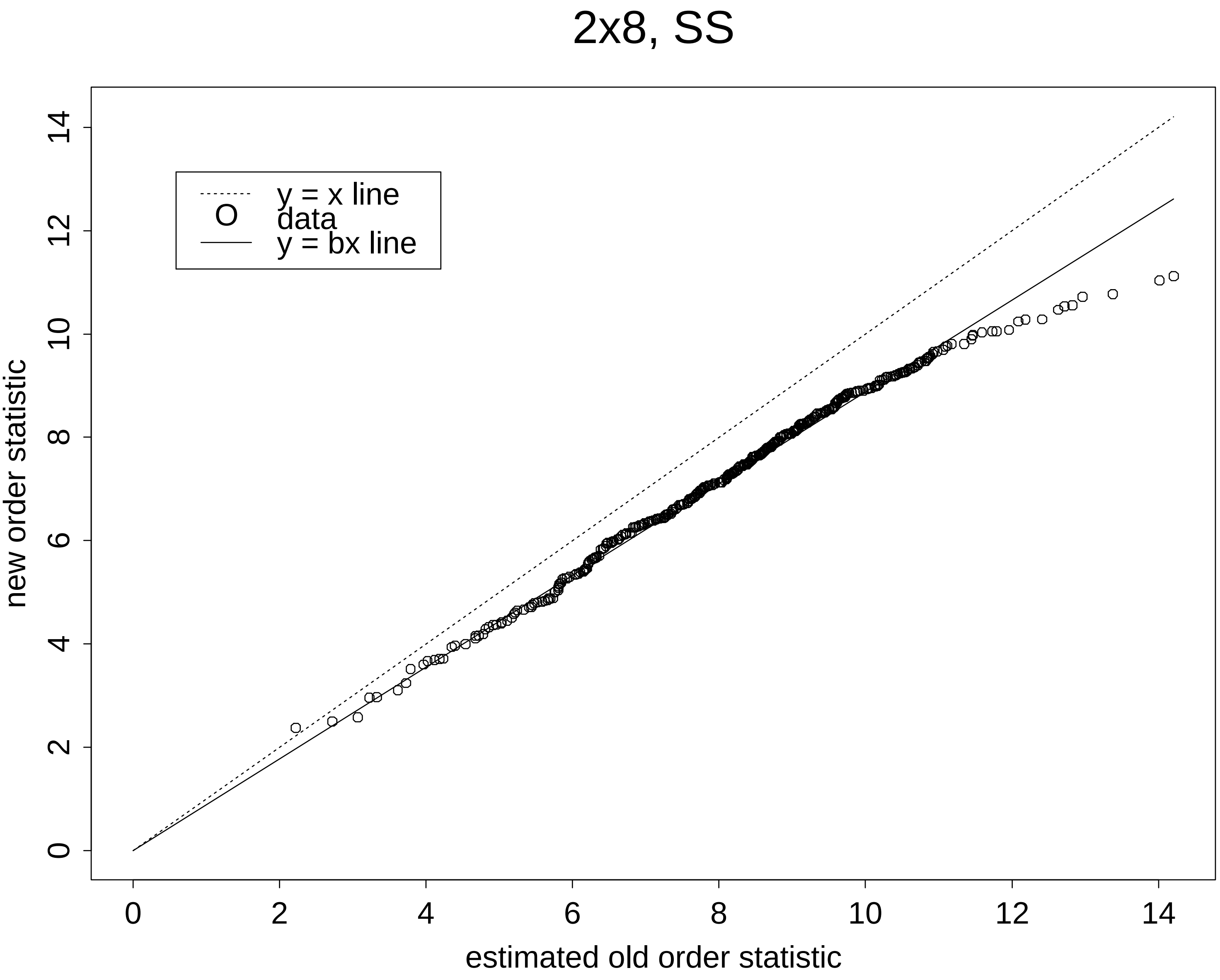

Figure 7: 2x8, SS data. Ordered monitoring data versus estimated ordered original data. The dotted line is the y = x line. The solid line is the fitted sortnew = b*oldest line.

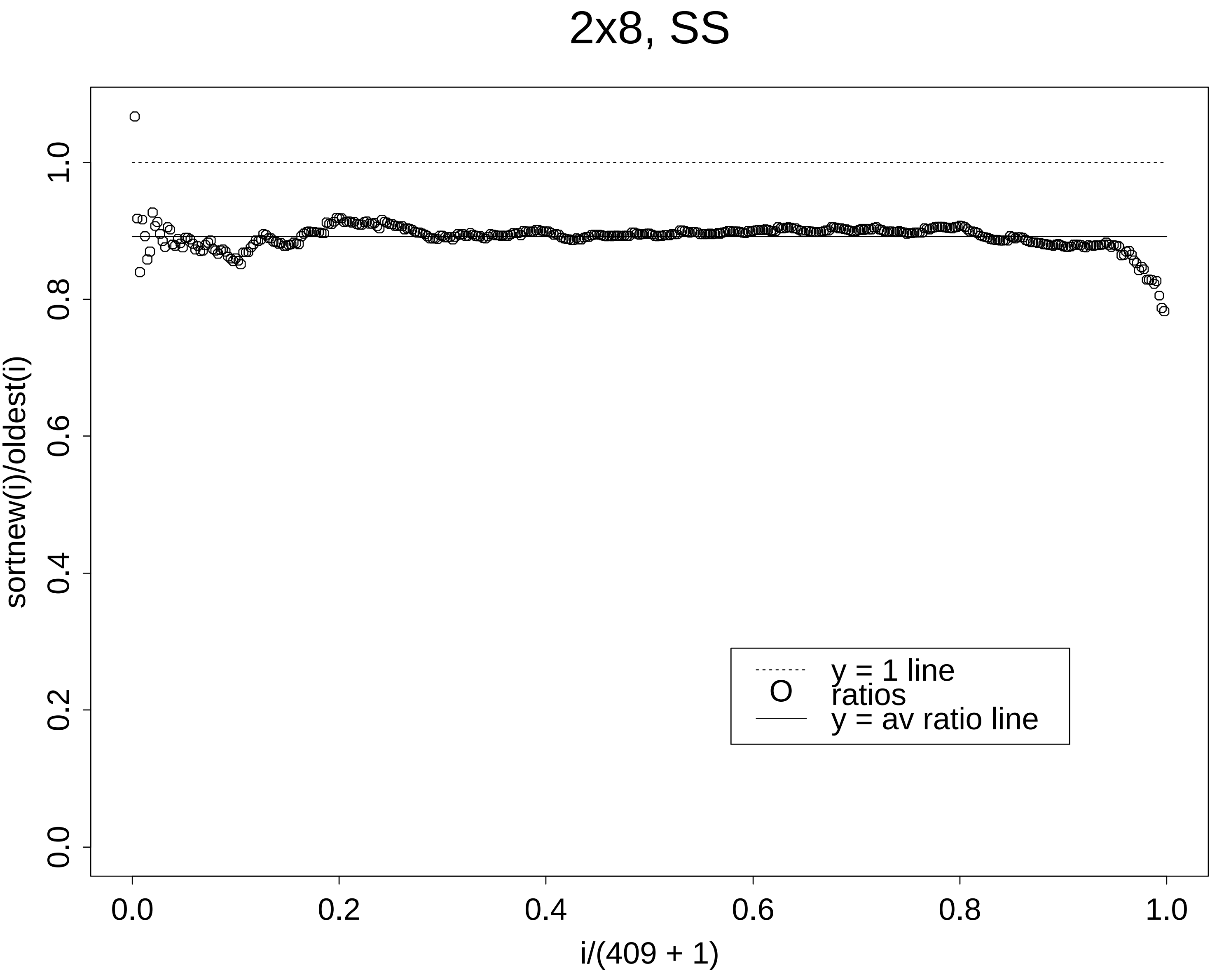

Figure 8: 2x8, SS data. sortnew(i)/oldest(i) versus $i/(409+1)$. The dotted line is the y = 1 line. The solid line is the y = average ratio line.

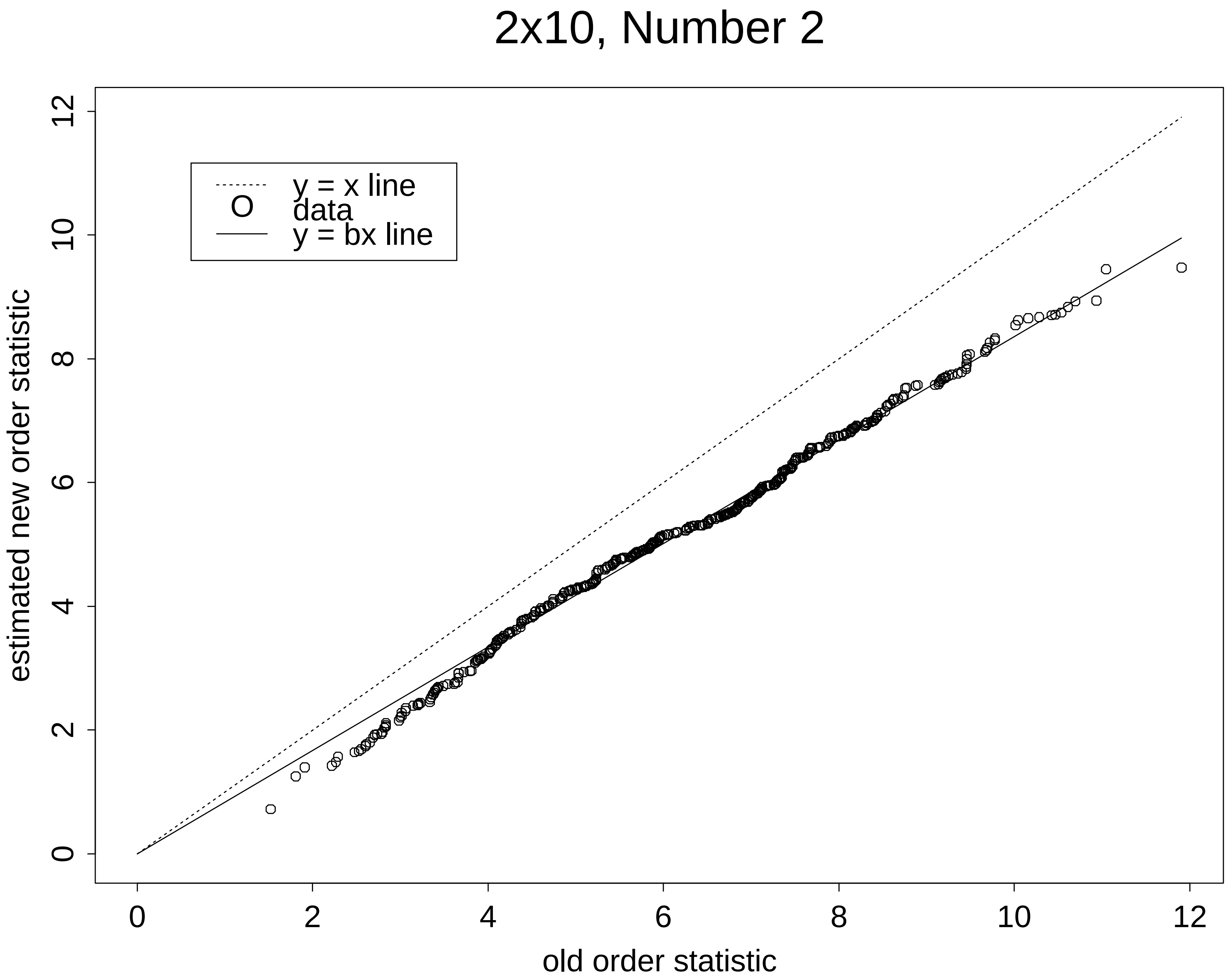

Figure 9: 2x10, Number 2 data. Estimated ordered monitoring data versus ordered original data. The dotted line is the y = x line. The solid line is the fitted newest = b*sortold line.

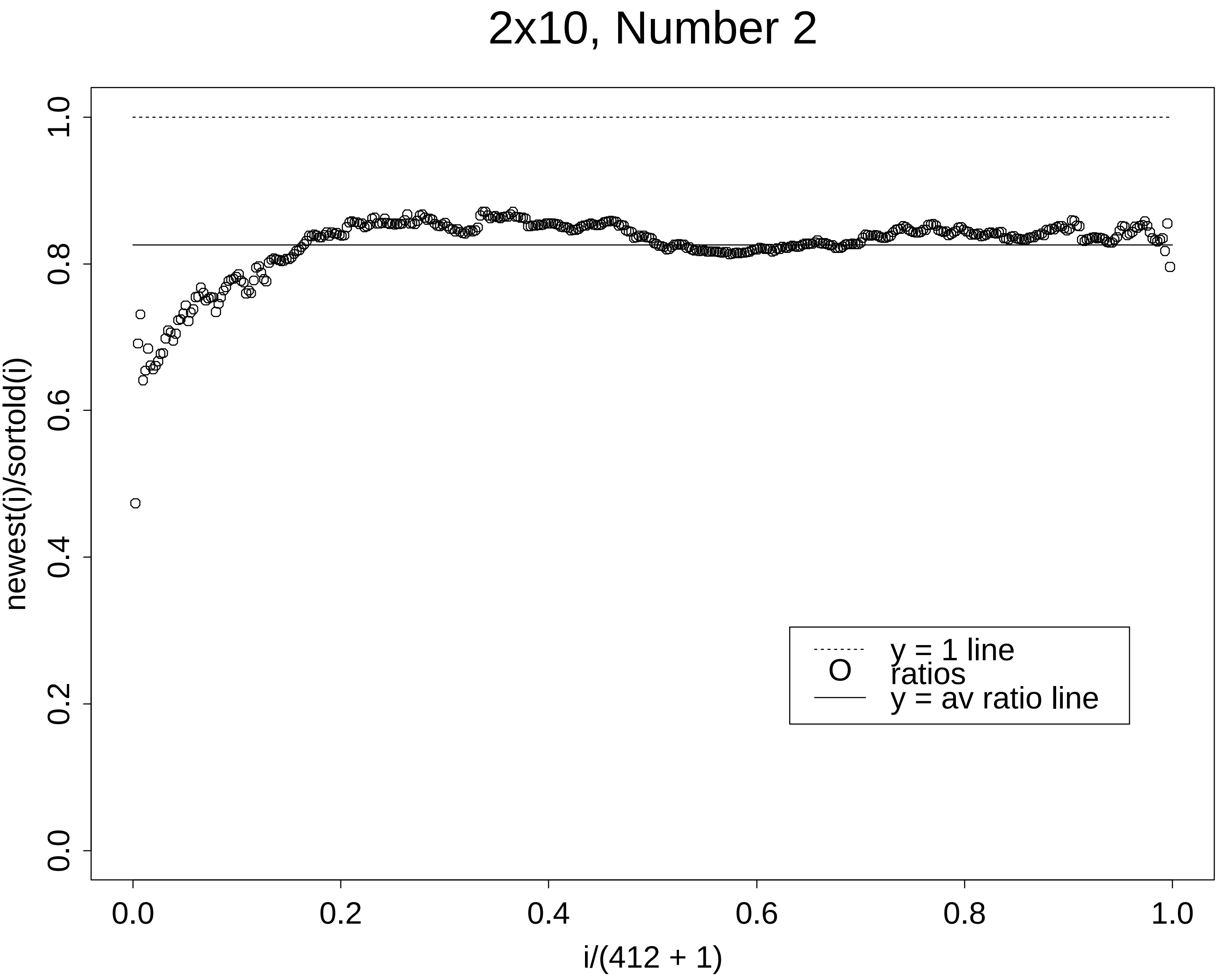

Figure 10: 2x10, Number 2 data. newest(i)/sortold(i) versus $i/(412 + 1)$. The dotted line is the y = 1 line. The solid line is the y = average ratio line.

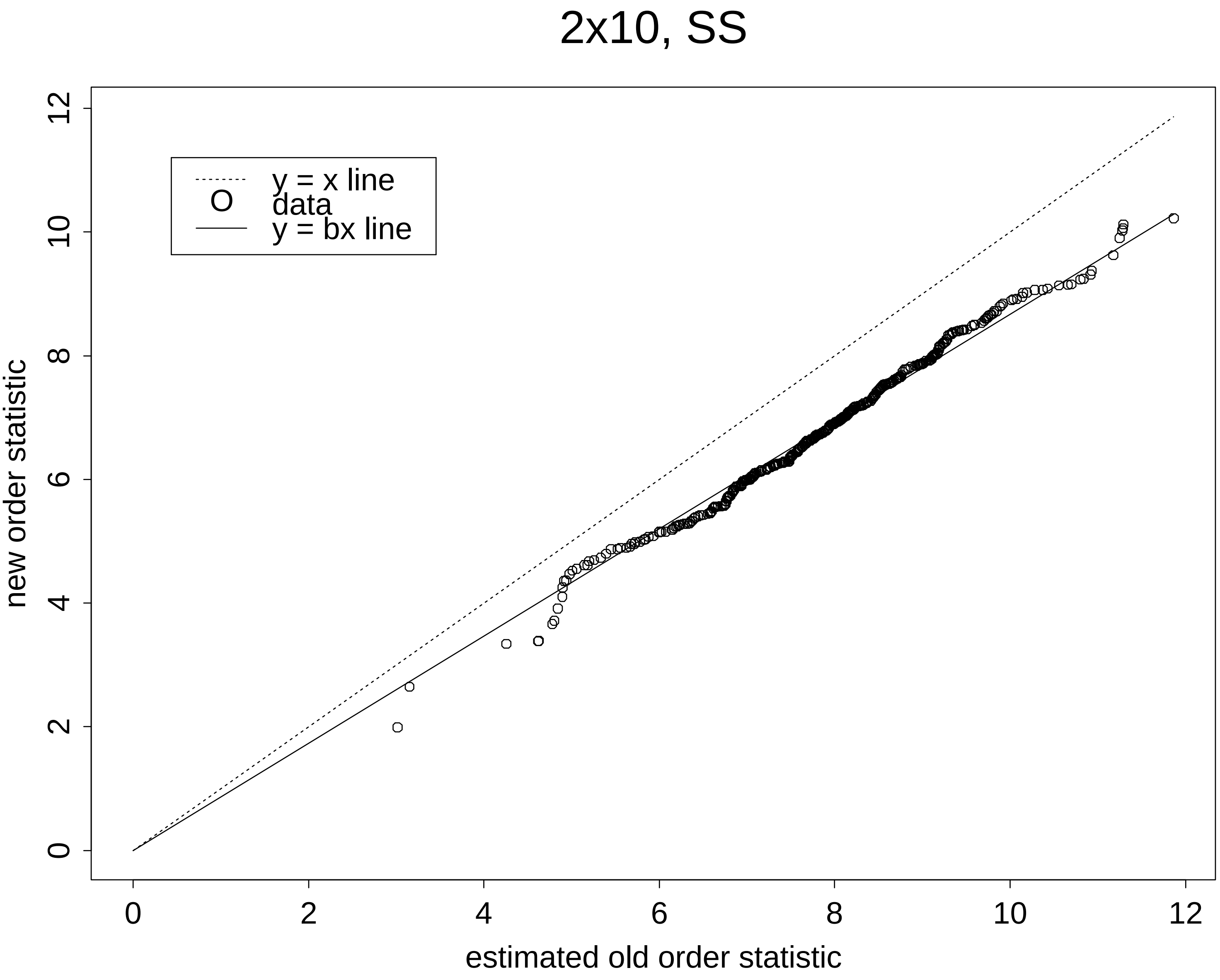

Figure 11: 2x10, SS data. Ordered monitoring data versus estimated ordered original data. The dotted line is the y = x line. The solid line is the fitted sortnew = b*oldest line.

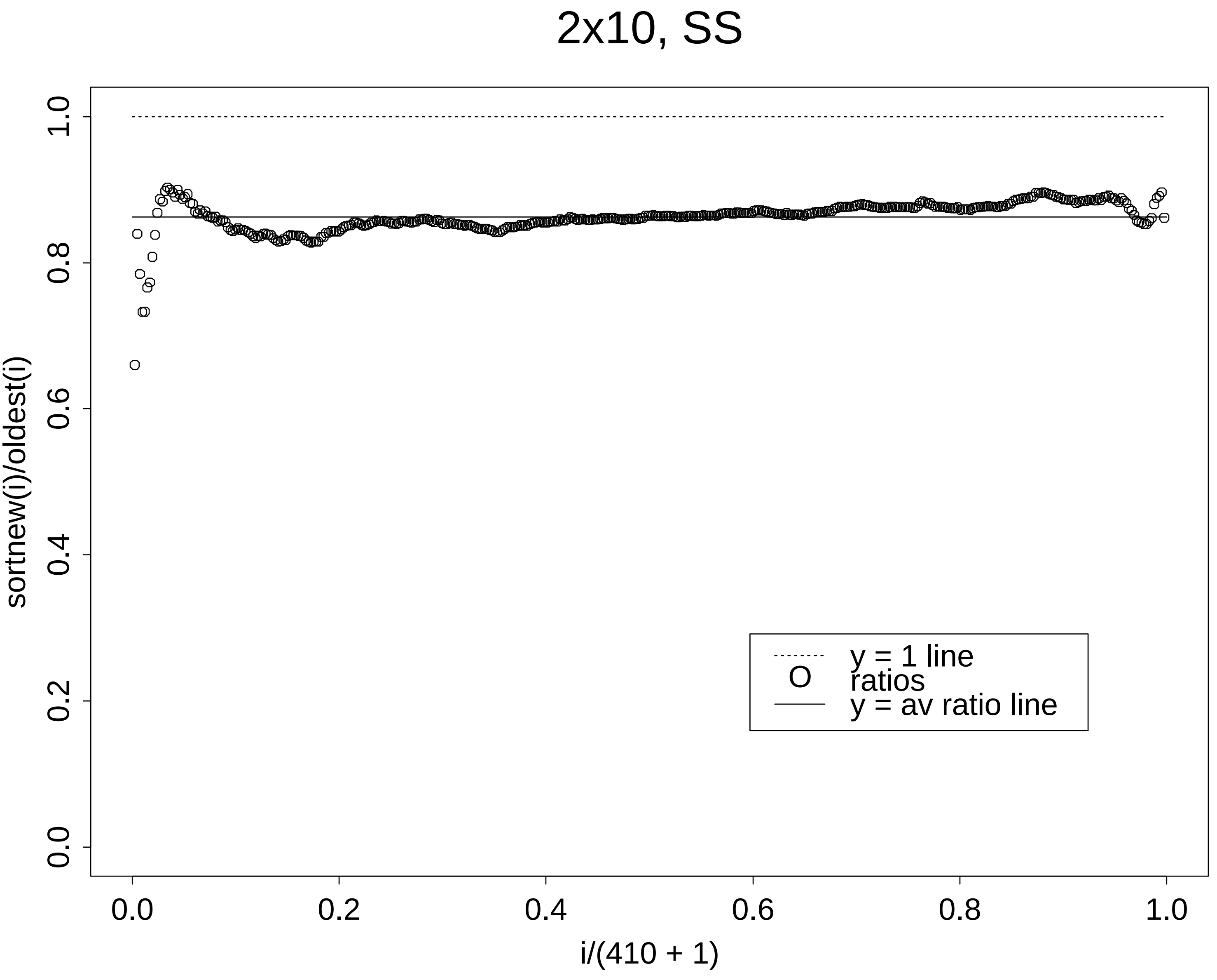

Figure 12: 2x10, SS data. sortnew(i)/oldest(i) versus $i/(410 + 1)$. The dotted line is the y = 1 line. The solid line is the y = average ratio line.

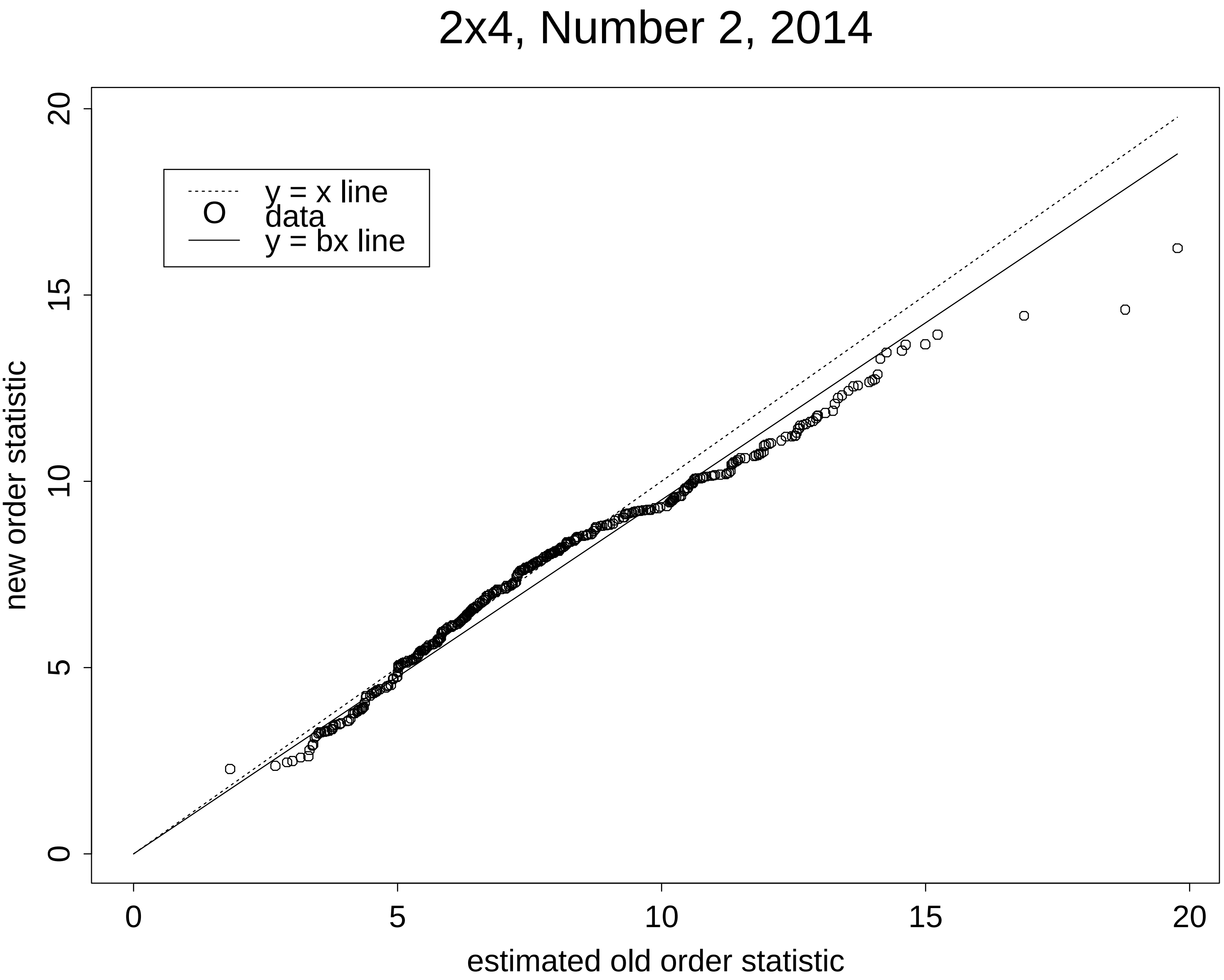

Figure 13: 2x4, Number 2 data, 2014. Ordered monitoring data versus estimated ordered original data. The dotted line is the y = x line. The solid line is the fitted sortnew = b*oldest line.

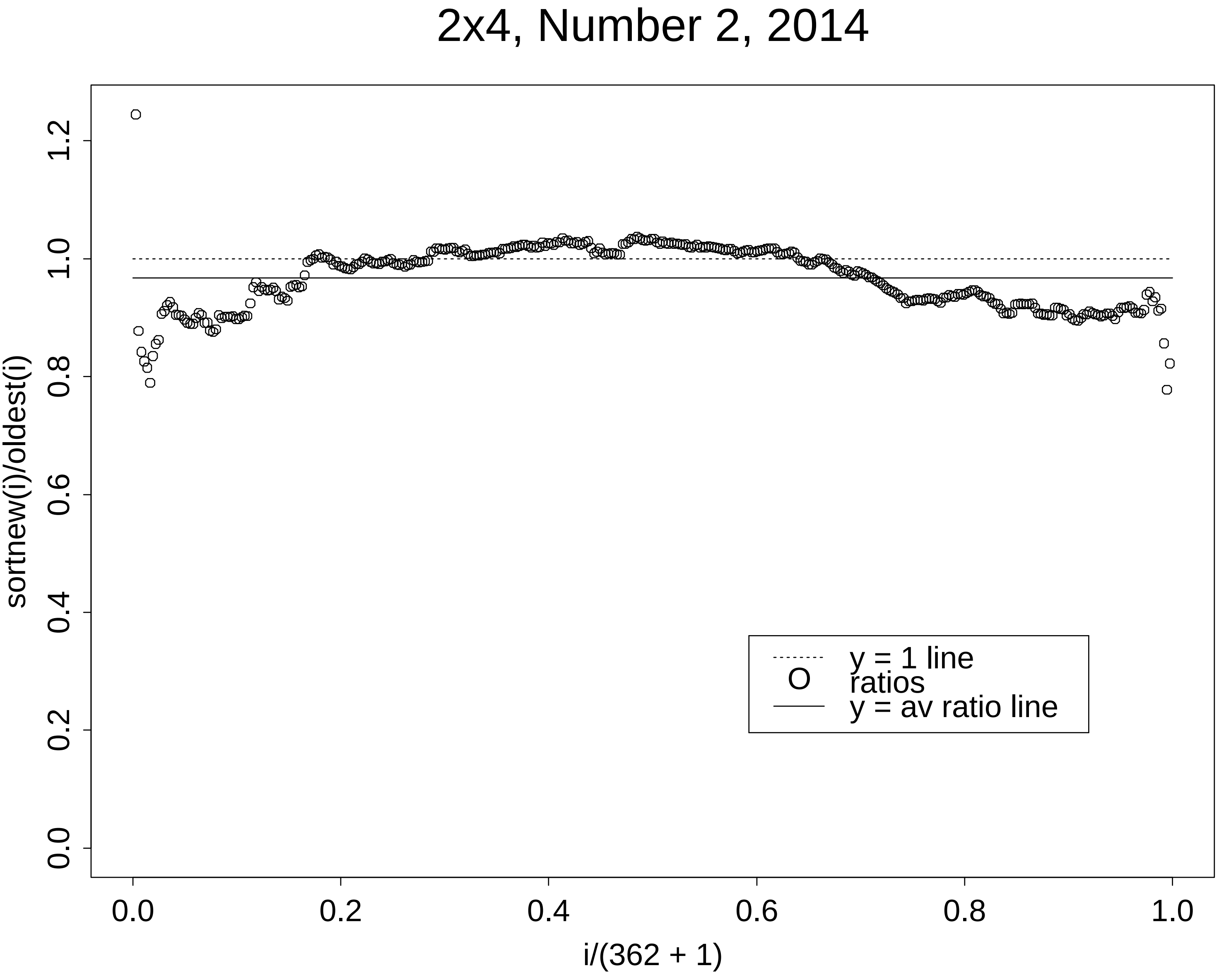

Figure 14: 2x4, Number 2 data, 2014. sortnew(i)/oldest(i) versus $i/(362+1)$. The dotted line is the y = 1 line. The solid line is the y = average ratio line. Note that the y-axis scale differs from the y-axis scales in Figures 2, 4, 6, 8, 10, and 12 (and the y-axis scale in Figure 8 differs from the scale in Figures 2, 4, 6, 10, and 12).