SIMULATIONS OF ERROR PROPAGATION FOR PRIORITIZING DATA ACCURACY IMPROVEMENT EFFORTS

(Research-in-progress)

Irit Askira Gelman
University of Arizona
Askirai@email.arizona.edu

Abstract: Models of the association between input accuracy and output accuracy imply that, for any given application, the effect of input errors on the output error rate generally varies in size depending on the choice of the specific input. While errors in one input may have a dramatic effect on the output error rate, a comparable or even higher error rate in another input may have a negligible effect. Clarification of this variation can be useful in data-quality management settings, since it can guide resource allocation decisions. Inputs in which errors exhibit a higher negative effect on the output would naturally earn higher priority.

The assistance that the models provide in the detection of such variation is, however, insufficient. Mainly, applying such a model can be painstaking. Therefore, there is a need to construct theories that illustrate the effects of errors in relatively broad scenarios. This study aims at such an introductory theory. The study applies simulations in order to illustrate error propagation in two basic information processing operations: the Boolean binary logical OR and logical AND. These operations are commonly used in the course of decision-making tasks. The results imply two simple rules for guiding data-management resource-allocation decisions. The findings also challenge common beliefs: they point to conditions in which a higher input accuracy generates lower output accuracy.

Keywords: Data Accuracy, Data Quality Management, Data Quality Improvement, Simulation, GIGO.

1. INTRODUCTION

Anecdotal evidence indicates that the cost of data errors can be astonishingly high. The overall cost of poor data quality to businesses in the US has been estimated at over $600 billion [12], and the overall cost to individual organizations is believed to be 10%-20% of their revenues [24]. However, such estimates are not impressive enough, apparently, to drive organizations to action—most of them do not have corporate data quality programs in place.

The disregard that organizations show for the quality of their data is explained by the general difficulty of assessing the economic consequences of the data quality factor, and the substantial cost that can be involved in achieving high data quality [12],[24]. The economic aspect of data quality has been drawing a growing research interest in recent years. An understanding of this aspect can be crucial for convincing organizations to address the data quality issue. It can guide decisions on how much to invest in data quality and how to allocate the limited organizational resources [28].

The economics of data quality, however, is partly determined by the relationship between the quality of the data and the quality of the information that the information system outputs. This is because data often undergo various processing before any actual use. An increasing number of studies have explored this relationship, mainly from a methodological perspective (e.g., [3], [5],[25],[22]), but also from empirical (e.g.,[17],[18]) and, rarely, analytical (e.g., [1]) perspectives. However, our grasp of the relationship

between an information system's input quality and its output quality is still often limited.

This study centers on the accuracy dimension of information quality. It is part of a research project that aims to clarify the association between input accuracy and output accuracy. The specific problem that this work addresses is explained below.

Models of the association between input accuracy and output accuracy (e.g.[3],[7]) imply that, for any given application, the effect of input errors on the output error rate generally varies in magnitude depending on the choice of the specific input. While errors in one input may have a dramatic effect on the output error rate, a comparable or even higher error rate in another input can have a negligible effect. Characterization of this variation is important since it can guide resource allocation decisions in data-quality management settings. The idea is to take into account the intended use of the data, such that inputs in which errors exhibit a higher negative effect on the output would earn higher priority. Nonetheless, the assistance that models like the above provide in this direction is insufficient. Although their application in specific instances can explain the variations in the effect of errors in those instances, applying such a model can be painstaking. Subsequently, in order to provide a simpler path to an insight on specific instances there is a need to analyze the effect of errors in relatively general scenarios.

This paper offers such initial insight through a series of simulations. The simulations designate two fundamental information processing operations: the Boolean binary logical OR and logical AND. These operations are commonly used in the course of decision-making [13]. Such decision processes can be described as follows. Given a set of decision variables, the values of matching dichotomous decision criteria are determined by testing the variable values against specified subsets of their domains. Subsequently, the values of the dichotomous criteria (a sequence of "True" and "False" values, for example) are combined using logical disjunction or conjunction to produce the outcome of the decision. The scenario assumed here is elementary. The association between the decision variables and respective dichotomous criteria is not considered. Instead, the inquiry focuses directly on effects of errors in the dichotomous criteria on the outcomes of disjunctive and conjunctive decisions that employ such criteria as inputs. An operation applies two inputs, both of which can contain errors. The correct input values as well as error occurrences are assumed random, and the relationship between input accuracy and output accuracy is interpreted as a relationship between the probability of input error occurrence and the probability of output error occurrence.

The outcome of the simulations implies two simple rules for guiding resource allocation decisions:

(1) When the percentage of correct values that satisfy the respective decision criterion varies across decision variables, efforts to improve the accuracy of the output of an OR operation should assign higher priority to errors in decision variables where a *higher* percentage of the correct values meet the matching criterion.

(2) When the percentage of correct values that satisfy the respective decision criterion varies across decision variables, efforts to improve the accuracy of the output of an AND operation should assign higher priority to errors in decision variables where a *lower* percentage of the correct values meet the matching criterion.

Importantly, the results of a related study that the author has conducted using mathematical-statistical methods clarify that these rules hold true in scenarios in which, given $N>2$ inputs, the output is derived through successive applications of the binary operation [1].

The simulations also reveal that the sign of the relationship between input accuracy and output accuracy can actually be negative. That is, contrary to the sweeping belief in the GIGO (Garbage In Garbage Out) assumption, higher input accuracy can indeed produce lower output accuracy. The change in the sign of the association follows a natural progression, as explained later.

The structure of this paper is as follows: Section 2 offers a brief summary of relevant literature. Section 3 describes the method in use by this study. Section 4 presents the results of the simulations. It portrays the variations in the magnitude and direction of the association between input accuracy and output accuracy under the Boolean binary logical OR and logical AND operations. Section 5 concludes the paper.

2. LITERATURE REVIEW

The problem of the relationship between an information system's input quality and its output quality has lately received much attention in the Data Quality (DQ) literature. For the most part, research has maintained a methodological nature.

Ballou and Pazer [3] proposed a framework for tracking numeric data errors through an information system, to assist with estimates of the impact of errors on the output. Their model takes into account errors due to both input inaccuracy and processing inaccuracy, though the emphasis is on how processing magnifies or dampens data errors. Ballou et al.[5] have applied and extended the model of Ballou and Pazer to other data quality dimensions. The extended model has been introduced together with a set of graphic modeling constructs that have been collectively called the Information Manufacturing Model (IMM). An IMM models an information system comparable to a data flow diagram (DFD). As a whole, Ballou et al.'s work enables the systematic tracking of timeliness, quality, and cost, and can be used to analyze an information system and assess various design alternatives from a data quality standpoint. Ballou et al.'s IMM has been augmented through the Information Product MAP (IPMAP) model [25] and subsequent enhancements of the IPMAP model. Research in this stream has drawn on inquiries in the accounting literature that offer decision aids for internal control evaluation based on mathematical modeling, or simulations. The proposed models generate overall system reliability or error rate estimates through aggregation of error rates in individual accounting processes (e.g., [27], [11], [29], [16]). Also related to this research stream are common models of error propagation that originated in the physical sciences literature [7].

Various frameworks have also been proposed in the context of database research for assessing the relationship between the quality of the raw data and the quality of the output of database queries [2],[21],[22]. A scenario that is more compatible with this work, especially its treatment of logical conjunction, is addressed by Ballou and Pazer [4]. Ballou and Pazer propose a framework for assessing the effect of input errors on the accuracy of decisions. They assume a dichotomous decision that is based on multiple criteria, such that the decision process applies a non-compensatory conjunctive rule for integrating the given criteria.

Empirical studies, often using simulations, have addressed the relationship between input accuracy and output accuracy assuming a diversity of prediction models (e.g., [17], [18]). These studies do not address the relative consequence of errors in distinct inputs but rather, they focus on the overall outcome.

The results of this study also contribute to a growing literature in various fields that is not entirely consistent with the belief in GIGO. A well-established theory in this category explains that statistical dependence relationships among data sources, or data errors, can have a dramatic effect on the accuracy of the information that an integration process produces (e.g., [6], [9], [10], [15], [19], [20]). This theory hints that higher data accuracy can lead to higher, or lower, output accuracy, subject to variations in statistical dependencies. Another relevant research stream includes studies of prediction model-building paradigms which indicate that adding noise to a data sample that serves in the construction of a model can improve the accuracy of the model (e.g., [8],[23],[26]). Evidently, controlled levels of noise can compensate for limitations of the model-building algorithms. That is, information-processing optimality seems to be a factor that can affect the sign of the link between input accuracy and output accuracy.

The inconsistency with GIGO that this work uncovers can not be attributed to statistical dependencies as above. Neither is it due to a sub-optimality of the chosen information processing operations. Nonetheless, the results can be explained by the nature of the Boolean binary AND and OR operations.

3. SIMULATION METHOD

The method employed here is Monte Carlo simulation. Monte Carlo simulation is a method for iteratively evaluating a deterministic model using sets of random numbers as inputs. The inputs are generated randomly from probability distributions to simulate the process of sampling from an actual population. Many simulations are then performed and the result is taken as an average over the number of data points in the sample [14].

The elements that comprise the simulation are described below.

Random numbers: The simulation process generates (pseudo) random instances of the following random variables:

- *U*, *V*: The ideal, correct input of the OR or AND operation; *U* and *V* are dichotomous random variables that accept the values 1 and 0, which correspond to *true* and *false*, respectively.
- D_U, D_V: Inform about the occurrence of an input error in the representation of *U* and *V*, respectively. These are dichotomous random variables that accept the values 1 and 0, which correspond to *error* and *no error*, respectively.

Instances of these variables are created from pre-determined probability distributions. Specifically, pre-determined collections of expected values of *U*, *V*, D_U, and D_V, denoted by p_U, p_V, p_{D_U}, and p_{D_V}, respectively, serve in the generation of random instances. Table 1 lists the expected values that we have used, as will be explained later. Notably, the expected value of a random variable that represents the occurrence of an error is the same as the probability of occurrence of that error.

Simulation model: Two deterministic models implement the information processing operations of interest, i.e., the Boolean binary logical OR and logical AND operations. These models are specified below (1)-(7). The models calculate p_{D_W} — the probability of an output error—based on a set of random values of *U*, *V*, D_U, and D_V.

The output of an error-free logical disjunction operation, *W*, is calculated using the equation:

$$W = U + V - UV \tag{1}$$

The consistency of (1) with the definition of logical disjunction can be easily verified through a systematic evaluation of *W* for each possible combination of the values of *U* and *V*. Similarly, the actual output, W_a, is derived from values of the actual inputs, U_a and V_a, according to the equation:

$$W_a = U_a + V_a - U_a V_a \tag{2}$$

where U_a and V_a in (2) are calculated from *U*, *V*, D_U, and D_V using:

$$U_a = (1 - D_U)U + D_U(1 - U) = U + D_U - 2UD_U \tag{3}$$

$$V_a = (1 - D_V)V + D_V(1 - V) = V + D_V - 2VD_V \tag{4}$$

If the value of D_U is zero, that is, if this variable indicates that no error has occurred, then (3) is reduced to $U_a=U$, i.e., the actual input is the same as the correct input. However, if the value of D_U indicates the occurrence of an error, then (3) assigns a value of one to U_a if *U* is zero and a value of zero if *U* is one. An equivalent relationship exists among V_a, D_V, and *V* (4).

Given *W* and W_a, the occurrence of an output error, denoted by D_W, is determined based on the relationship among W_a, D_W, and *W* , which is comparable to (3) and (4):

$$W_a = (1 - D_W)W + D_W(1 - W) = W + D_W - 2WD_W \quad (5)$$

Finally, the probability of an output error, p_{D_W}, is estimated by the average of D_W over the number of generated sample instances.

In the case of logical conjunction, the output of an error-free conjunction operation, W, is calculated through:

$$W = UV \quad (6)$$

The consistency of (6) with the definition of logical conjunction can be verified by a systematic evaluation of W for each possible combination of the values of U and V. Analogously, the actual output, W_a, is derived from the actual inputs, U_a and V_a, according to the equation:

$$W_a = U_a V_a \quad (7)$$

The computations of the values of the available inputs, U_a and V_a, apply (3) and (4), again. The occurrence of an output error, D_W, is determined using (5). As in the case of logical disjunction, the probability of an output error is estimated by averaging the value of D_W over the number of generated sample instances.

Sample size: Each simulation applies a sample of 90,000 instances of each of U, V, D_U, and D_V. Given this sample size, we can derive an estimate of the error of the simulation result. Let σ_{D_W} denote the standard deviation of D_W. It is easy to see that, under the assumptions of this investigation:

$$\sigma_{D_W}^2 = p_{D_W} - p_{D_W}^2 \quad (8)$$

Since the standard deviation of the outcome of a Monte Carlo analysis decreases with the square root of the sample size [14], we conclude that σ, the standard deviation of our estimate of p_{D_W}, satisfies:

$$\sigma = \frac{\sigma_{D_W}}{\sqrt{90,000}} = \frac{\sigma_{D_W}}{300} = \frac{\sqrt{p_{D_W} - p_{D_W}^2}}{300} < \frac{\sqrt{p_{D_W}}}{300} \quad (9)$$

Consequently, if, for example, p_{D_W} is in the range 0.01-0.15, then σ is in the range 0.0003-0.0012, respectively.

p_U	p_V	p_{D_U}	p_{D_V}	Total no. of simulations
0.01, 0.03, 0.05,...,0.99 (50 different values)	p_U, p_U +0.02, p_U +0.04,..., 0.99	0.01,0.02,...,0.15 (15 different values)	0.01,0.02,...,0.15 (15 different values)	$\frac{(50+1)\cdot 50}{2} \cdot 15 \cdot 15$ =286,875

Table 1: Number of simulations and implemented parameter values.

Number of simulations and implemented parameter values: Altogether, 286,875 simulations were carried out. Each simulation generated a sample matching a distinct combination of p_U, p_V, p_{D_U}, and

p_{D_V} . The combinations that were implemented by the simulations are recorded by Table 1. Each of the specified values of every parameter was implemented in combination with every registered value of any of the other parameters.

The simulations were performed by GAUSS, a mathematical-statistical programming language.

4. RESULTS

The results of the simulations of the OR operation are portrayed in Figure 1. The results of the simulations of the AND operation are portrayed in Figure 2 in the Appendix. Due to space limitations, Figure 1 and Figure 2 depict a representative selection of the results rather than the entire result set. Each graph refers to a unique choice of the means of the correct values of the inputs, p_U and p_V, and describes the output error rate, p_{D_W}, as a function of the input error rates p_{D_U} and p_{D_V}, in the range 0.01-0.15. The values of p_U and p_V are specified by the title of each graph. For example, the three graphs at the top of Figure 1 describe the output error rate as a function of the input error rates when p_U =0.01 and p_V = 0.01, 0.11, and 0.21, respectively. The values of p_U and p_V were selected such that p_U varies in the range 0.01-0.99, and for each value of p_U, p_V varies between p_U and 0.99. Since the inputs are symmetric, this set nearly covers the range of all possible values of the means.

A review of Figure 1 and Figure 2 reveals the following patterns:

1. The means of the correct values, p_U and p_V, affect the output error rates. Graphs vary depending on the choice of p_U and p_V.
2. As the average of p_U and p_V grows higher, the output of an OR operation (Figure 1) is *less* sensitive to input errors, while the output of an AND operation (Figure 2) is *more* sensitive to input errors.
3. For the most part, the surfaces that depict the results are not symmetric with respect to the diagonal plane p_{D_U} = p_{D_V} . Additional study of Figure 1 indicates that under the OR operation, if the means of the correct values vary significantly then the output is generally more susceptible to errors in the input whose correct values have the *higher* mean. Figure 2 shows that under the AND operation, if the means of the correct values vary significantly then the output is more susceptible to errors in the input whose correct values have the *lower* mean.
4. Some of the graphs uncover a striking behavior, namely, as input error rates increase, output error-rates decrease (note the declining slopes). Such a negative association is demonstrated when the means of the correct values are extremely different. For example, a negative association is observed in the graph that refers to p_U =0.01 and p_V =0.99, and the graph that refers to p_U =0.01 and p_V =0.95.

Next, we discuss the above observations in detail given each of the information processing operations.

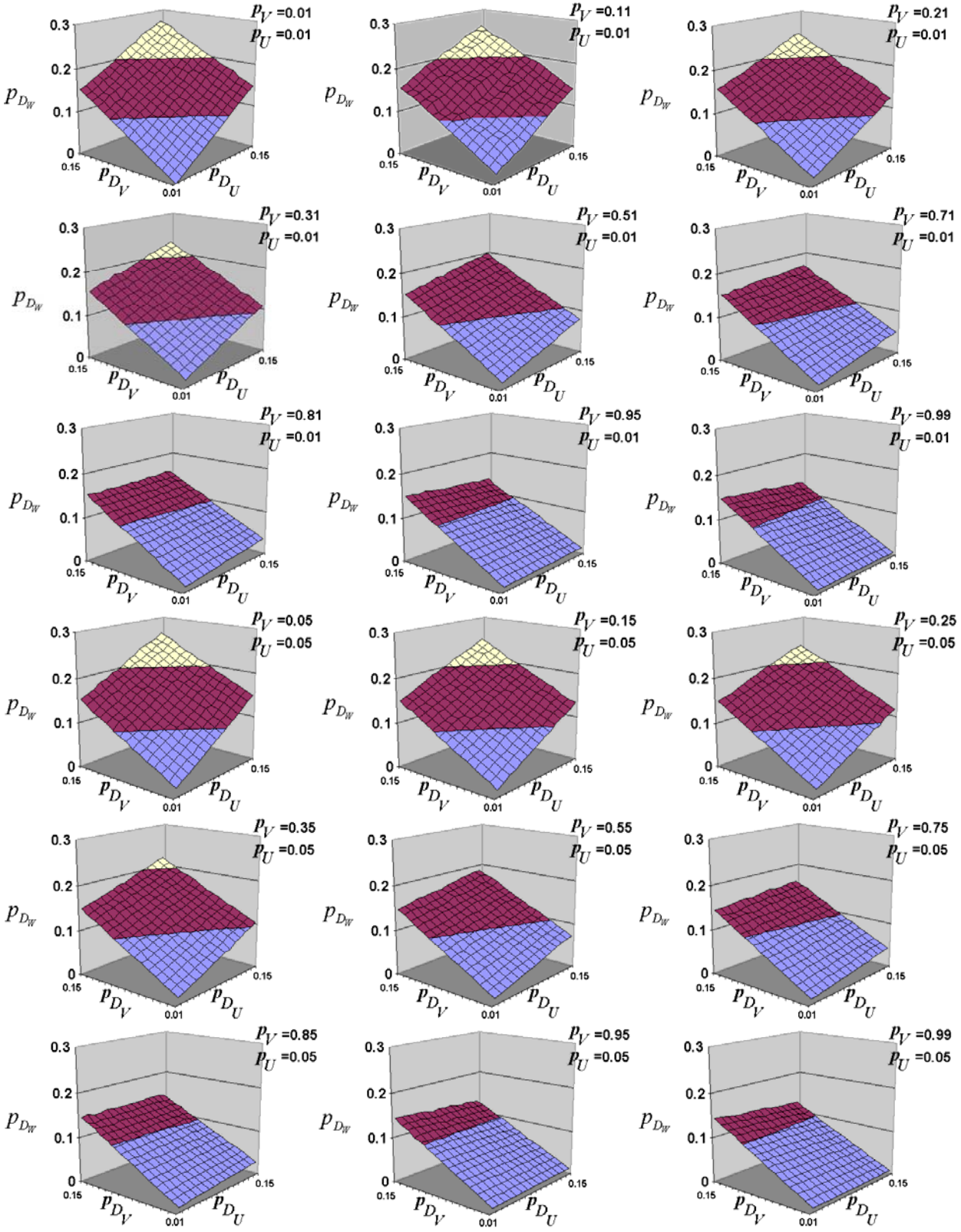

Figure 1: Simulations of error propagation under the OR operation.

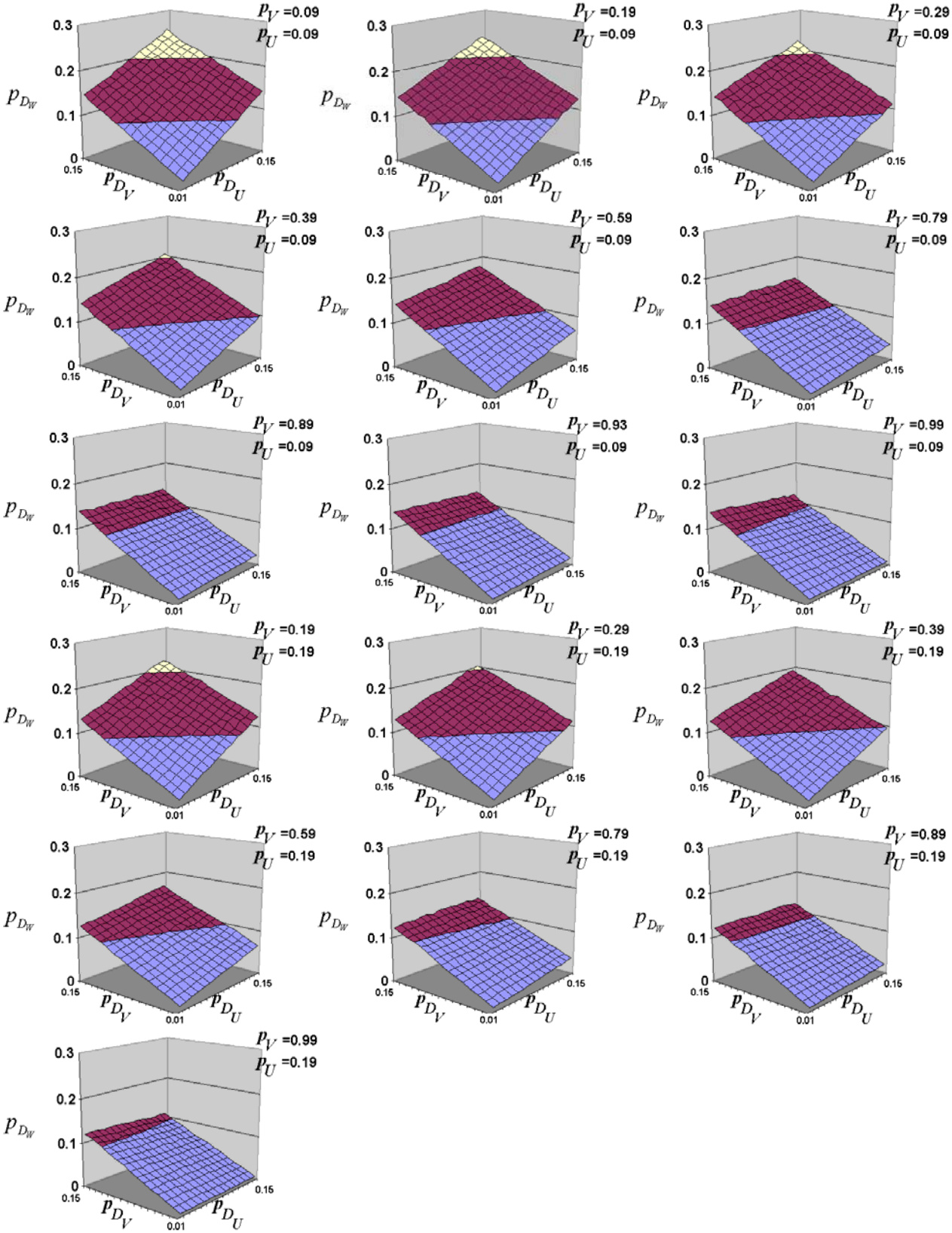

Figure 1: Simulations of error propagation under the OR operation. (CONTINUED)

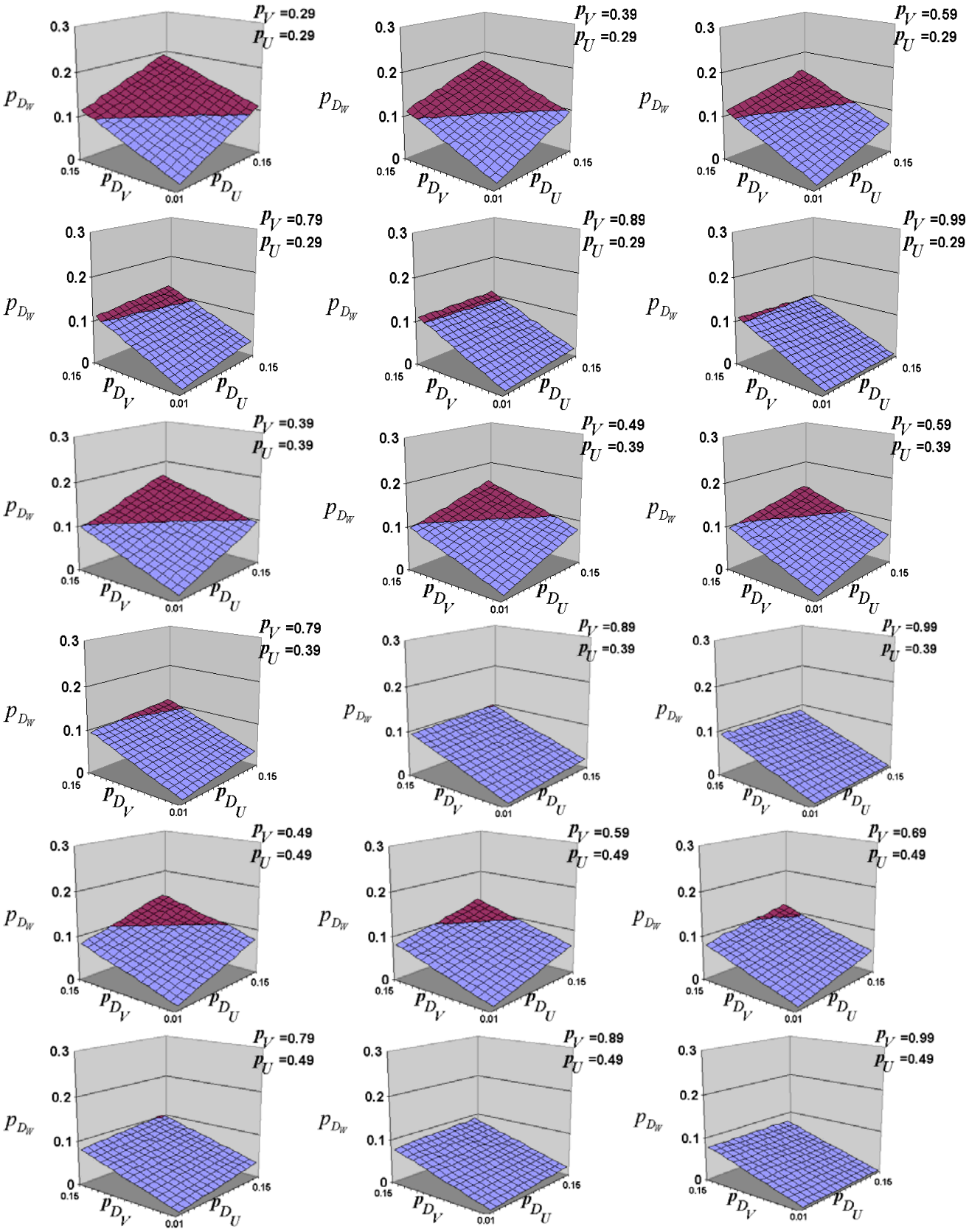

Figure 1: Simulations of error propagation under the OR operation. (CONTINUED)

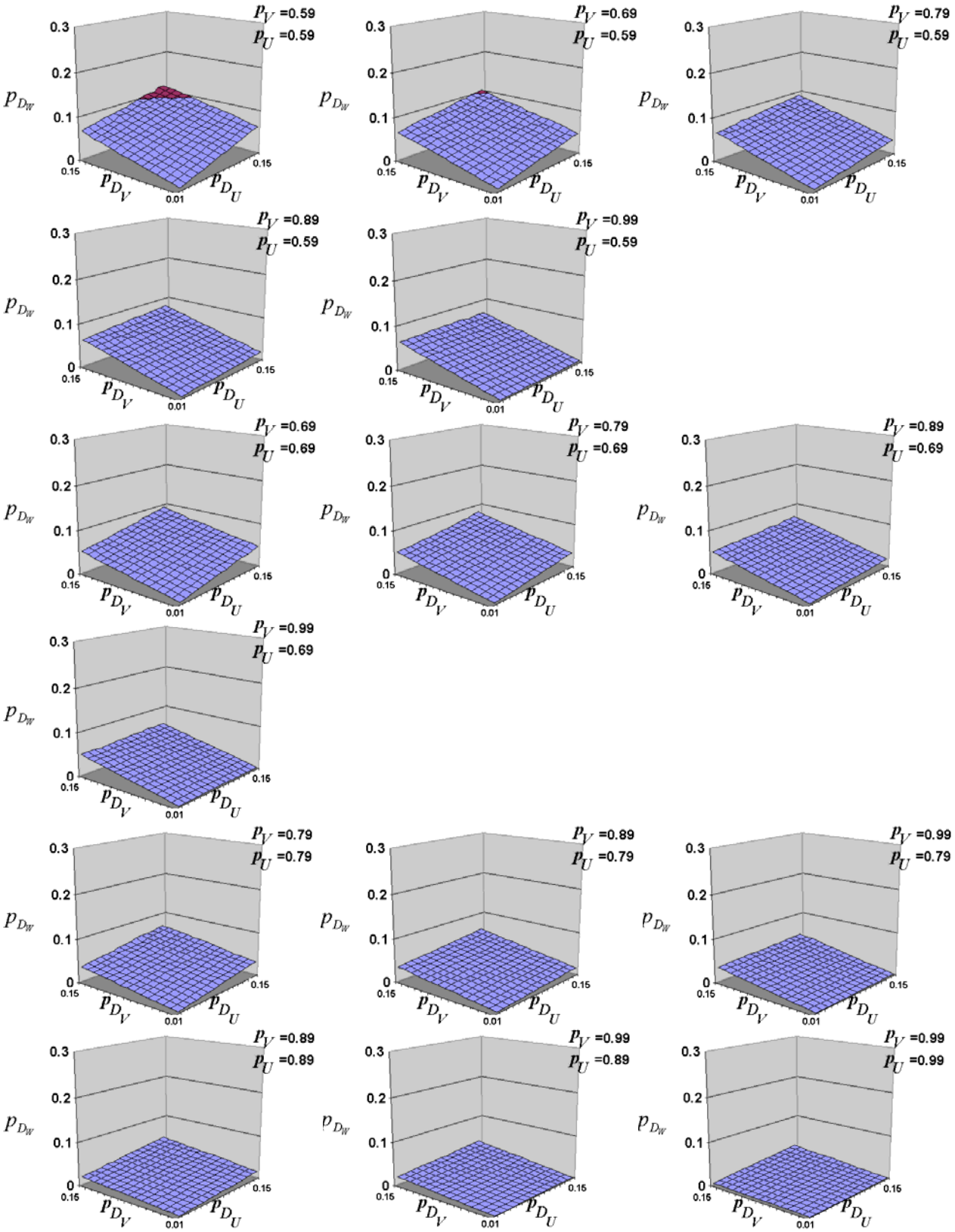

Figure 1: Simulations of error propagation under the OR operation. (CONTINUED)

4.1 Results: OR

The susceptibility of the output of an OR operation to input errors decreases as the average mean of the correct values grows higher. Intuitively, the output of an OR operation is most affected by input errors when the correct values of the inputs are equal to zero. In this case, an error in any of the inputs induces an output error since the output changes from zero to one. The output of an OR operation is least affected by input errors when the correct values of the inputs are equal to one. In this case, a single error in any of the inputs does not generate an output error at all. However, as the average of p_U and p_V increases, the probability of a combination of zeros declines while the probability of a combination of ones rises.

When the means of the correct values vary across inputs, the output is more sensitive to errors in an input with a higher mean. Graphs that are asymmetric with respect to the diagonal plane $p_{D_U} = p_{D_V}$ make up the majority in Figure 1. A symmetry is evident only in the graphs that designate p_U and p_V such that $p_U = p_V$; in all other cases, the graphs are asymmetric. This suggests that the effect of a given input error rate on the output error rate varies depending on the input that exhibits such error rate. However, a deeper study of Figure 1 suggests a far stronger conclusion. *If the means of the correct values are sufficiently far apart from each then the output is invariably more sensitive to errors in the input with the higher mean.* (Under the assumptions of this investigation, the former statement is valid for any p_U and p_V such that $p_V > p_U + 0.14$, although for many values of p_U and p_V a smaller gap is enough.)

Here is how such a conclusion is derived from Figure 1. Setting aside fluctuations due to the limited accuracy of the simulations, every graph in Figure 1 describes a surface which, for any fixed value of p_{D_V}, is reduced to a straight line.[1] The slope of such a line, which reflects the change in the output error rate as p_{D_U} varies, depends on p_{D_V} as well as on p_U and p_V. The graphs demonstrate that the slope varies gradually in its magnitude as p_{D_V} increases from p_{D_V}=0.01 to p_{D_V}=0.15, such that it is always at its extrema at p_{D_V}=0.15 and p_{D_V}=0.01. Similarly, for any fixed value of p_{D_U}, the surface is reduced to a straight line with comparable characteristics. Now, when we review the graphs in Figure 1 and, for each surface, compare the slopes of the lines p_{D_V}=0.15 and p_{D_V}=0.01 with the slopes of the lines p_{D_U}=0.15 and p_{D_U}=0.01, we reach an interesting conclusion. In all the graphs such that p_V is sufficiently higher than p_U (e.g., $p_V > p_U + 0.14$), the magnitudes of the slopes at p_{D_U}=0.15 and p_{D_U}=0.01 are *both* higher than the magnitudes of the slopes at p_{D_V}=0.15 and p_{D_V}=0.01. Therefore, in these conditions the output error rate is systematically affected more by changes in p_{D_V} than by variations in p_{D_U}, regardless of the values of these error rates. Plainly, if the means of the correct values are sufficiently far apart from each other then the output is invariably more sensitive to errors in the input with the higher mean. It can be proved that the required gap between those means is, at most, greater than the maximal possible gap between the input error rates, $|p_{D_V} - p_{D_U}|$.

An intuitive explanation attributes the unequal importance of the inputs to the relatively high frequency of instances in which the correct value of the input with the higher mean is one while the correct value of the input with the lower mean is zero. If, given such a combination of correct values, an error occurs in the value one, then the OR operation would produce an error (i.e., the output would be zero instead of one). In contrast, an error in the input where the correct value is zero would not affect the result of the OR operation.

[1] This statement is based on a mathematical –statistical analysis [1].

Negative association between the input error rate and the output error rate. A negative association between p_{D_U} and p_{D_W} is observed in graphs that designate p_U and p_V such that p_V is extremely high relative to p_U. A negative association is a natural continuation of a weakening positive effect. For any p_U, the positive association between p_{D_U} and p_{D_W} that is exhibited when $p_V = p_U$ becomes weaker as p_V increases. Ultimately, when p_V is exceptionally high compared to p_U, the positive association between p_{D_U} and p_{D_W} changes to a negative association. That is, errors in the input whose correct values have the lower mean eventually demonstrate a negative association. A negative association reflects a situation in which errors in the input with the lower mean have the role of "good errors." That is, they offset the "bad errors" in the input with the higher mean.

4.2 Results: AND

The susceptibility of the output of an AND operation to input errors increases as the average mean of the correct values increases. The output of an AND operation is most affected by input errors when the correct values of the inputs are equal to one, and is least affected by input errors when the correct values of the inputs are equal to zero. However, as the average of p_U and p_V increases the probability of a combination of zeros declines while the probability of a combination of ones rises.

When the means of the correct values vary across inputs, the output is more sensitive to errors in an input with a lower mean. Figure 2 can be analyzed similar to Figure 1 to demonstrate that when the means of the correct values are sufficiently far apart from each other (again, $p_V > p_U + 0.14$ although a smaller gap would often be satisfactory) then the output is invariably more sensitive to errors in the input with the *lower* mean.

As before, our explanation links the unequal importance of the inputs to the relatively high frequency of instances in which the correct value of the input with the higher mean is one while the correct value of the input with the lower mean is zero. If, given such combination of correct values, an error occurs in a zero, then the AND operation would produce an error (i.e., the output would be one instead of zero). In contrast, an error in the input where the correct value is one would not affect the result of the AND operation.

Negative association between the input error rate and the output error rate. A negative association is, again, a natural continuation of a weakening positive effect. For any p_U, the positive association between p_{D_V} and p_{D_W} that is shown when $p_V = p_U$ grows weaker as p_V increases. Eventually, when p_V is especially high compared to p_U, the positive association between p_{D_V} and p_{D_W} changes to a negative association. In other words, errors in the input whose correct values have the higher mean ultimately demonstrate a negative association. A negative association matches a situation in which errors in the input with the higher mean have the role of "good errors," which offset the "bad errors" in the input with the lower mean.

5. CONCLUSIONS

Decision processes that entail dichotomous decision criteria are widespread. Given a set of decision variables, the values of corresponding dichotomous decision criteria are determined by testing the variables against specified subsets of their domains. Often, the outcome of a decision is produced by combining the values of the dichotomous criteria through logical disjunction or conjunction. This paper

examined the effect of errors in dichotomous decision criteria on the output of such decisions, based on the observation that the effect of errors in different inputs on the output error rate generally varies in magnitude. The viewpoint that drives this study is that such variation can be useful in data-quality management settings since it can guide resource allocation decisions— inputs in which errors display a higher negative effect on the output would gain higher priority. When resources are limited, an ability to establish priorities while taking into account the intended use of the data can be valuable.

The findings of the simulations imply two simple rules for guiding resource allocation decisions:

(1) When the percentage of correct values that satisfy the respective decision criterion varies across decision variables, efforts to improve the accuracy of the output of an OR operation should assign higher priority to errors in decision variables where a higher percentage of the correct values meet the matching criterion.

(2) When the percentage of correct values that satisfy the respective decision criterion varies across decision variables, efforts to improve the accuracy of the output of an AND operation should assign higher priority to errors in decision variables where a lower percentage of the correct values meet the matching criterion.

According to a related study that the author has conducted using mathematical-statistical methods, these rules are valid also in scenarios in which, given $N>2$ inputs, the output is derived through successive applications of the binary operation [1]. Notably, as the two resource allocation rules reveal, implementation of these rules requires, for each decision criterion, an estimate of the percentage of the correct values that meet the criterion. As for error rates, typically there would be no need to study them much in advance. An observation that error rates are not zero combined with rough estimates of the maximum boundaries of the error rates of the decision variables would be sufficient. Situations in which the percentages of the correct values that meet the criteria are approximately equal should be clarified by future research.

The results also suggest that when the percentage of correct values that satisfy the corresponding decision criterion varies radically across decision variables, the sign of the relationship between input accuracy and output accuracy can be negative. This finding challenges the strong belief in the GIGO assumption. Overall, however, our understanding of the relationship between an information system's input accuracy and its output accuracy is lacking. Therefore, future work should continue the pursuit of this important relationship.

REFERENCES

1. Askira Gelman, I., "Setting Priorities in Data Accuracy Improvement Projects: A Guiding Theory." unpublished manuscript.
2. Avenali, A., Bertolazzi, P., Batini, C., and Missier, P., "A Formulation of the Data Quality Optimization Problem in Cooperative Information Systems." *International Workshop on Data and Information Quality in conjunction with CAISE'04*, Riga, Latvia, 2004.
3. Ballou, D. P. and Pazer, H. L., "Modeling Data and Process Quality in Multi-input, Multi-output Information Systems." *Management Science*, Vol. 31, No. 2, 1985, pp. 150-162
4. Ballou, D. P. and Pazer, H. L., "A Framework for the Analysis of Error in Conjunctive, Multi-Criteria, Satisficing Decision Processes." *Decision Sciences*, Vol. 21, No. 4, 1990, pp. 752-770.
5. Ballou, D. P., Wang, R. Y., Pazer, H. L., and Tayi, G. K., "Modeling Information Manufacturing Systems to Determine Information Product Quality." *Management Science*, Vol. 44, No. 4, 1998, pp. 462-484.
6. Barabash, T. L. "On properties of symbol recognition." *Engineering Cybernetics*, 1965, pp. 71-77.
7. Bevington, P. R. *Data Reduction and Error Analysis for the Physical Sciences.* New York: McGraw-Hill, pp. 58-64, 1969.

8. Bishop, C.M. “Training with noise is equivalent to Tikhonov regularization.” *Neural Computation*, 7(1), 1995, pp. 108-116.
9. Clemen, R.T., and Winkler, R.L. “Limits for the precision and value of information from dependent sources.” *Operations Res.*, 33(2), 1985, pp. 427-442.
10. Cover, T. “The best two independent measurements are not the two best.” *IEEE Transactions on Systems, Man and Cybernetics*, Vol. SMC-4, No. 1, 1974, pp. 116-117.
11. Cushing, B. E. “A mathematical approach to the analysis and design of internal control systems.” *Accounting Review*, January 1974, pp. 24-41.
12. Eckerson, Wayne W. "Achieving Business Success through a Commitment to High Quality Data," TDWI Report Series, The Data Warehousing Institute, 2002, p. 5.
13. Einhorn, H.J. (1970). The use of nonlinear, noncompensatory models in decision making. *Psychological Bulletin*, 73(3), pp. 221-230.
14. Fishman, G.S., *Monte Carlo: Concepts, Algorithms, and Applications*, Springer Verlag, New York, 1995.
15. Frantsuz, A.G. “Influence of correlations between attributes on their informativeness for pattern recognition.” *Engineering Cybernetics*, No 4, 1967.
16. Hamlen, S. S. “A chance-constrained mixed integer programming model for internal control design.” *Accounting Review*, October 1980, pp. 578-93.
17. Klein, B.D., and Rossin, D.F., “Data Quality in Linear Regression Models: Effect of Errors in Test Data and Errors in Training Data on Predictive Accuracy.” *Informing Science*, Vol. 2, No. 2, 1999a.
18. Klein, B.D., and Rossin, D.F., “Data Quality in Neural Network Models: Effect of Error Rate and Magnitude of Error on Predictive Accuracy.” *Omega*, Vol. 27, No. 5, 1999b, pp. 569-582
19. Kuncheva, L.I., Whitaker, C.J., Shipp, C.A., and Duin, R.P.W. “Limits on the majority vote accuracy in classifier fusion.” *Pattern Analysis and Applications*, 6(1), 2003, pp. 2-31.
20. Ladha, K. “Information pooling through majority-rule voting: Condorcet’s Jury Theorem with correlated votes.” *J. Econ. Behavior and Organization*, Vol. 26, 1995, pp. 353-372.
21. Motro, A., and Rakov, I. “Not All Answers Are Equally Good: Estimating the Quality of Database Answers.” In *Flexible Query-Answering Systems* (T. Andreasen, H. Christiansen, and H.L. Larsen, Editors). Kluwer Academic Publishers, 1997, pp. 1-21.
22. Parssian. A., Sarkar, S., and Varghese, S.J.,”Assessing Data Quality for Information Products: Impact of Selection, Projection, and Cartesian Product.” *Management Science*, Vol. 50, No. 7, 2004, pp. 967-982.
23. Raviv, Y., and Intrator, N. “Bootstrapping with noise: An effective regularization technique.” *Connection Science*, Special issue on Combining Estimators, 8, 1996, pp. 356-372.
24. Redman, T.C., “Data: An Unfolding Disaster.” *DM Review Magazine*, August 2004.
25. Shankaranarayan, G., Zaid M., and Wang, R., ”Managing Data Quality in Dynamic Decision Environments: An Information Product Approach. ” *Journal of Database Management*, Vol. 14, No. 4, 2003.
26. Skurichina, M., Raudys, S., and Duin, R.P.W. “K-Nearest neighbours directed noise injection in multilayer perceptron training. “ *IEEE Transactions on Neural Networks*, 11(2), 2000, pp. 504-511.
27. Stratton, W.O. “Accounting Systems: The reliability Approach to Internal Control Evaluation.” *Decision Sciences*, Vol. 12, No. 1, 1981, pp. 51-67.
28. Wang, R.Y., and Strong, D.M., "Beyond Accuracy: What Data Quality Means to Data Consumer." *Journal of Management Information Systems*, 12, 1996, pp. 5-34.
29. Yu, S. and J. Neter. “A stochastic model of the internal control system.” *Journal of Accounting Research*, Autumn 1973, pp. 273-295.

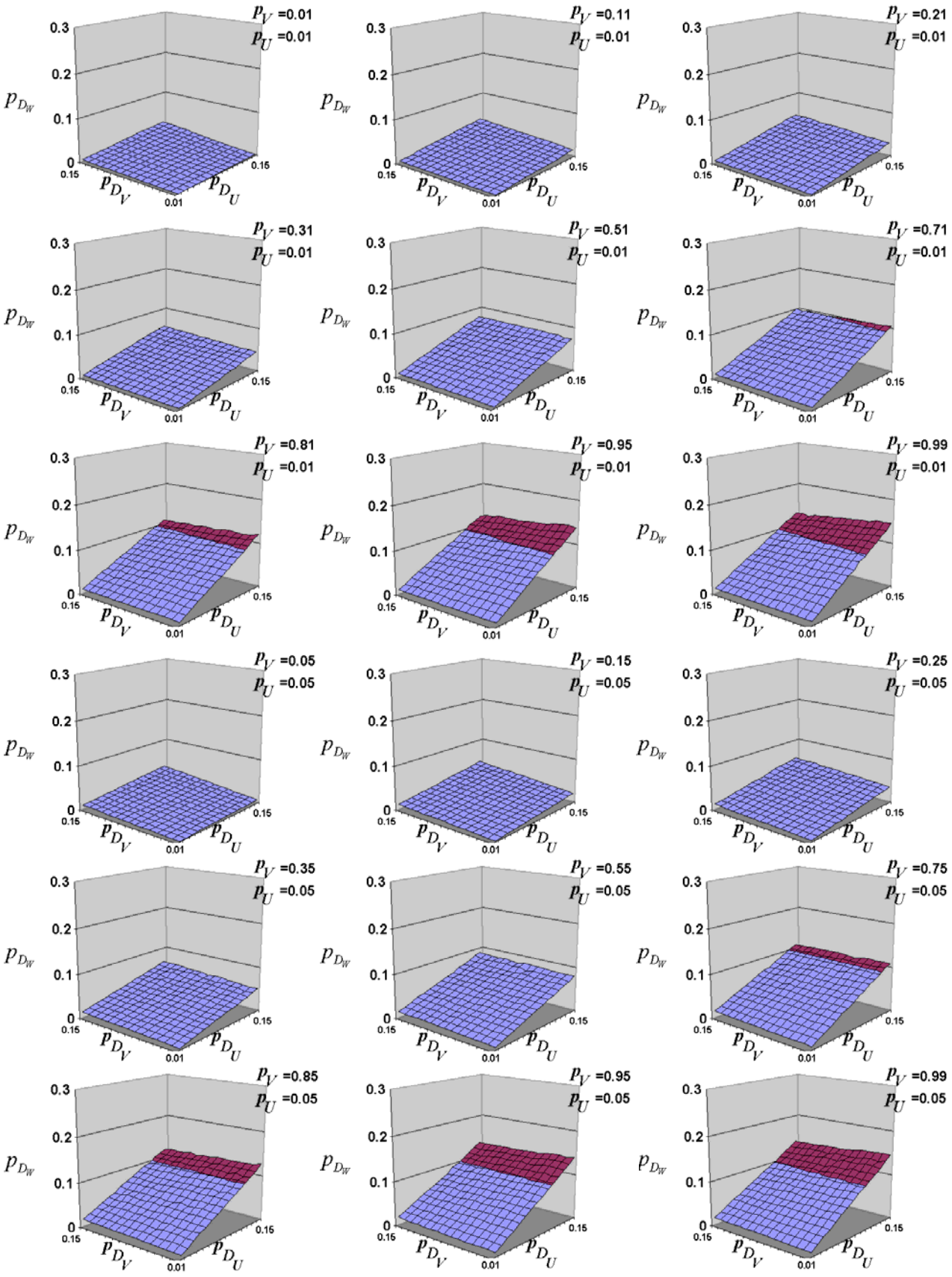

Figure 2: Simulations of error propagation under the AND operation.

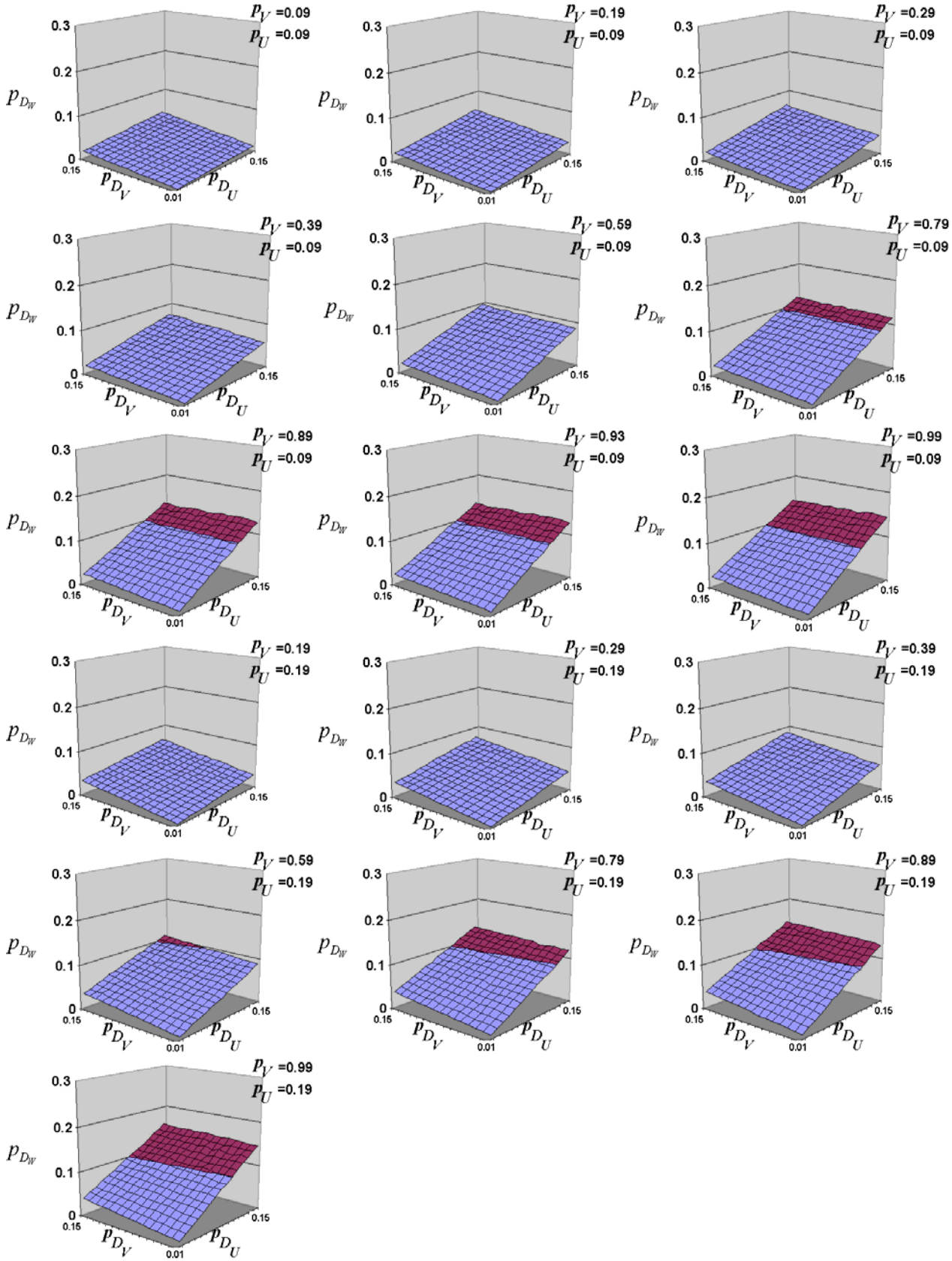

Figure 2: Simulations of error propagation under the AND operation. (CONTINUED)

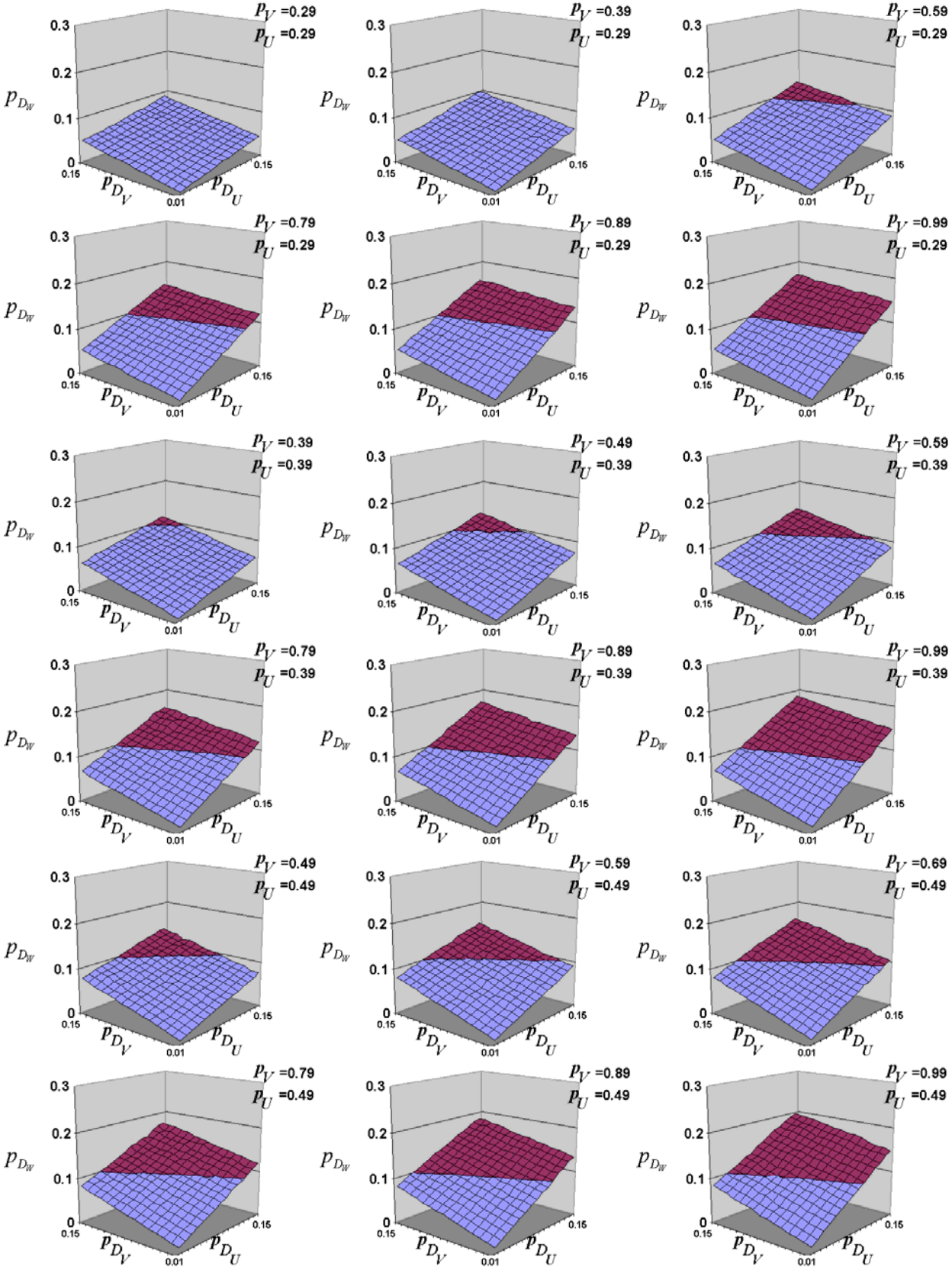

Figure 2: Simulations of error propagation under the AND operation. (CONTINUED)

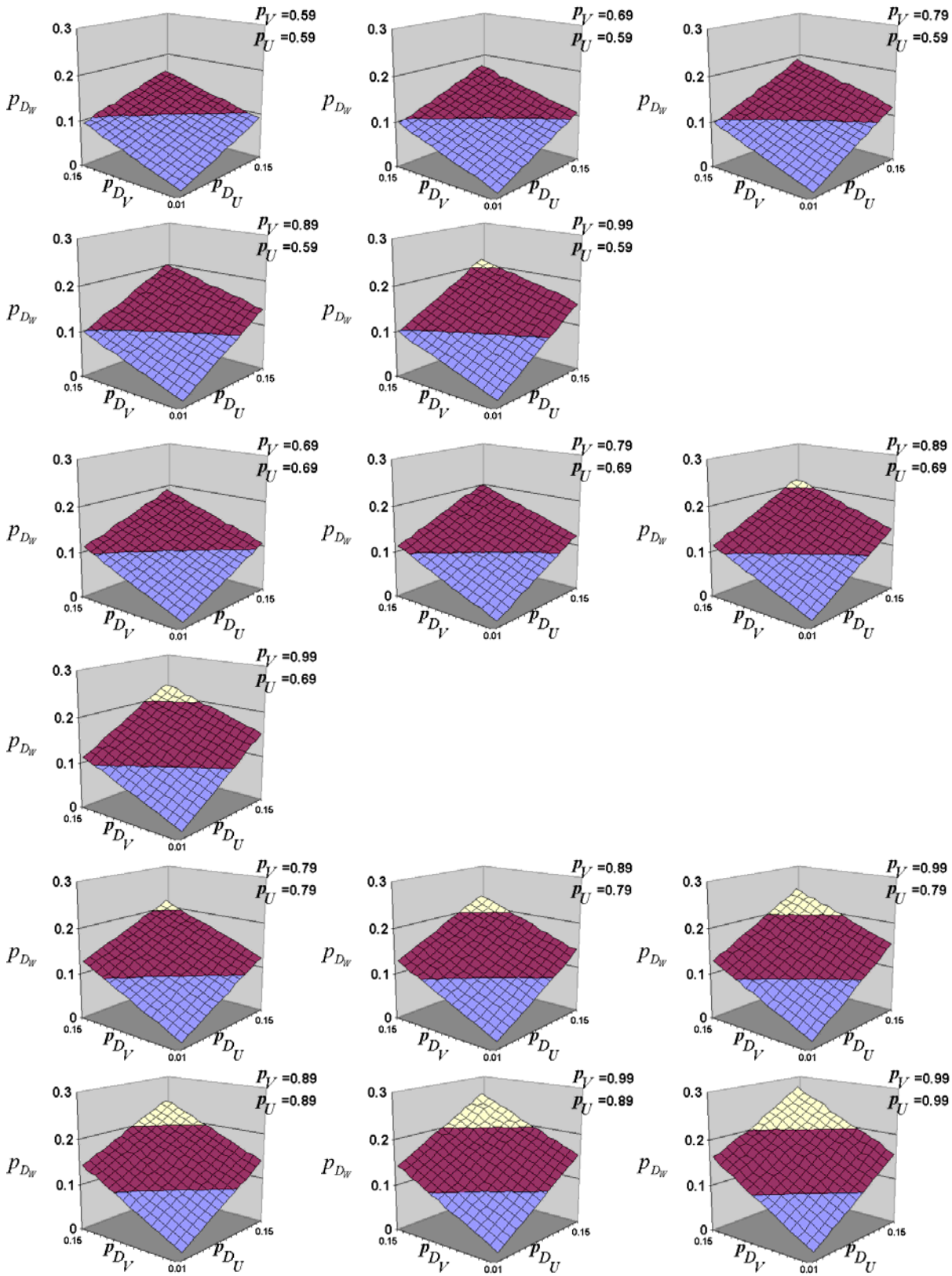

Figure 2: Simulations of error propagation under the AND operation. (CONTINUED)