Chapter 14
An Application of Exploratory Data Analysis in the Development of Game-Based Assessments

Kristen E. DiCerbo, Maria Bertling, Shonté Stephenson, Yue Jia, Robert J. Mislevy, Malcolm Bauer, and G. Tanner Jackson

Abstract While the richness of data from games holds promise for making inferences about players' knowledge, skills, and attributes (KSAs), standard methods for scoring and analysis do not exist. A key to serious game analytics that measure player KSAs is the identification of player actions that can serve as evidence in scoring models. While game-based assessments may be designed with hypotheses about this evidence, the open nature of game play requires exploration of records of player actions to understand the data obtained and to generate new hypotheses. This chapter demonstrates the use of the 4R's of Exploratory Data Analysis (EDA): revelation, resistance, re-expression, and residuals to gain close familiarity with data, avoid being fooled, and uncover unexpected patterns. The interactive and iterative nature of EDA allows for the generation of hypotheses about the processes that generated the

K.E. DiCerbo (✉)
Pearson, 400 Center Ridge Dr, Austin, TX 78753, USA
e-mail: kristen.dicerbo@pearson.com

M. Bertling • Y. Jia
Educational Testing Service, MS-T-03, 660 Rosedale Rd, Princeton, NJ 08541, USA
e-mail: mbertling@ets.org; yjia@ets.org

S. Stephenson
GlassLab Games, 209 Redwood Shores Pkwy, Redwood City, CA 94065, USA
e-mail: shonte.berkeley@gmail.com

R.J. Mislevy
Educational Testing Service, Center for Advanced Psychometrics,
MS 12-T, 660 Rosedale Rd, Princeton, NJ 08541, USA
e-mail: rmislevy@ets.org

M. Bauer
Educational Testing Service, 16-R, Turnbull Hall, Rosedale Rd, Princeton, NJ 08540, USA
e-mail: mbauer@ets.org

G.T. Jackson
Educational Testing Service, 660 Rosedale Rd, MS 16-R, Princeton, NJ 08541, USA
e-mail: gtjackson@ets.org

C.S. Loh et al. (eds.), *Serious Games Analytics*, Advances in Game-Based Learning, DOI 10.1007/978-3-319-05834-4_14

observed data. Through this framework, possible evidence pieces emerge and the chapter concludes with an explanation of how these can be combined in a measurement model using Bayesian Networks.

Keywords Exploratory data analysis • Game-based assessment • Evidence model • Data visualization • Re-expression • Residuals

1 Introduction

The past decade has seen a growing push for games in learning spaces (Gee, 2003). A new generation of promising educational games has emerged allowing for deep exploration of broad concepts (Klopfer, Osterweil, & Salen, 2009). Games support sociocultural and situative approaches to learning in which players interact with peers and their environment to develop knowledge and understanding of the world (Steinkuehler, 2004). In addition, data from games provide information about the process a player used to arrive at a final product, suggesting great potential for generating new insights regarding student actions as they relate to complex knowledge, skills, and attributes (Mislevy, Behrens, DiCerbo, Frezzo, & West, 2012). Game-based assessments (GBAs) have the potential to combine the rich problems, engagement, and motivation from games with the evidentiary arguments of assessment.

However, the potential of games as assessment tools can be met only if replicable methods for aligning game play with learning standards and formative assessment objectives can be developed. New interactive digital games elevate both the availability of student micro-patterns (small, repeatable segments of play actions) and the importance of understanding them as they reflect variation in strategy or evolving psychological states. While the richness of the data holds promise for making important inferences, standard methods for scoring and analysis do not exist. In addition, the open nature of many games means students often engage in unexpected actions in the game. This requires multiple cycles of data exploration, hypothesis generation, and confirmation on the part of the analyst to fully understand the relationships of game play actions to inferences about players.

Assessment is fundamentally about designing situations which elicit evidence about aspects of what learners know and can do. Evidence-Centered Design (ECD; Mislevy, Steinberg, & Almond, 2002) provides a framework for specifying these arguments. It defines the following models:

- Student model—What we want to know about the learner
- Task model—What activities the learner will undertake
- Evidence model—How we link the work produced in the task to the constructs in the student model. The evidence model contains two pieces:
- Scoring model—How we will identify evidence in the learners' work product
- Measurement model—The statistical techniques we use to link the evidence to the elements in the student model

This chapter will focus largely on the scoring model, or the identification of the important elements in the record of player actions to extract and pass to our measurement models. For multiple choice items, the scoring model is simple. The work product is a list of selected options. The scoring rule for each item is, "if selection matches correct response, then mark correct, otherwise mark incorrect." However, when the work product is a log file of actions a student has taken in a game, it is less clear how to identify the scoring rules, much less apply them. What are the actions in the game that will tell us about the knowledge, skills, and attributes of interest? Our usual assessment routines and psychometric processes cannot be easily lifted from our traditional assessments and applied to GBAs.

In designing GBAs, the specification of the scoring model is an iterative process. Design begins with hypotheses about what player actions will be important for making inferences. However, most games are complex systems. Before diving directly into confirming these hypotheses, it is important we understand the data obtained from the game and also seek to uncover unexpected patterns in the data that may generate new hypotheses. Exploratory Data Analysis (EDA; Tukey, 1977) provides a helpful framework by which to consider the processes of hypothesis generation and exposition of patterns in data. While EDA techniques are not new, the application of these older (but often overlooked) methods in this new context provides a way to facilitate new ways of identifying evidence for inferences about player knowledge, skills, and attributes. This chapter will focus on the use of EDA to gain close familiarity with game-based assessment data, avoid being fooled, and uncover unexpected patterns while developing an understanding of what features of player game play provide evidence about our constructs of interest. The final section of the chapter will demonstrate how these uncovered evidence fragments can then be inserted into a measurement model to estimate proficiency of game players. The scoring model and measurement model in combination allow the translation of game play into inferences about knowledge, skills, and attributes. The chapter will use analysis of data from SimCityEDU to demonstrate the concepts of the EDA framework.

1.1 Exploratory Data Analysis

EDA is a conceptual framework with a core set of ideas and values aimed at providing insight into data, and to encourage understanding probabilistic and nonprobabilistic models in a way that guards against erroneous conclusions (Behrens, DiCerbo, Yel, & Levy, 2012). EDA also provides a set of tools that allow researchers to become intimately familiar with their data. It encourages the development of mental models of the data and processes that created them.

EDA holds several complementary goals: to find the unexpected, avoid being fooled, and develop rich descriptions. The primary analogy used by Tukey (1977) to communicate these goals is that of the data analyst as detective. The work is essentially exploratory and interactive, involving an iterative process of generating

hypotheses and looking for fit between facts and the tentative theory or theories. Detective work also provides a solid analogy for EDA because both are essentially bottom-up processes of hypothesis formulation and data collection.

Tukey (e.g., 1986) did not consider methodology as a bifurcation between exploratory and confirmatory, but considered quantitative methods to be applied in stages of exploratory, rough confirmatory, and confirmatory data analyses. In this view, EDA is aimed at the initial goals of hypothesis generation and pattern detection following the detective analogy. It is therefore differentiated from the (correctly) maligned practice of snooping through data to find the data and model that will most likely lead to significant results. Rather, EDA generates hypotheses that are later confirmed with separate data. Rough confirmatory data analysis is sometimes equated with null-hypothesis significance testing that is often what is taught in statistics courses. Strict confirmatory analyses involve the more sophisticated testing of specific relationships and contrasts that is less common in research practice. As a researcher moves through these stages, she moves from hypothesis generation to hypothesis testing and from pattern identification to pattern confirmation.

In the context of EDA, the data analyst performs an iterative series of interactions with the data, all the while generating various observations and hypotheses about the forms of the data and the likely underlying processes that generated them. Therefore, to return to the original problem, EDA allows us to iteratively generate hypotheses about the patterns in the game data and their relationships to levels of knowledge, skills, and attributes. EDA provides a set of tools by which to accomplish this. We can think of them in relation to four R's (Hoaglin, Mosteller, & Tukey, 1983): revelation, re-expression, resistance, and residuals. Revelation refers to uncovering the unexpected, largely through visualization. Re-expression involves careful understanding of the distributions of variables. Resistance implies using methods that are not overly influenced by extreme or unusual data. Finally, residuals provide a means by which to evaluate and iterate with models. Each of these will be discussed further with examples in the remainder of the chapter.

1.2 Context

For illustrative purposes, references will be made throughout the chapter to SimCityEDU (www.simcityedu.org), developed by GlassLab. SimCityEDU, based on the popular SimCity commercial game, offers players various challenges that ask players to solve problems facing a city, generally requiring them to balance elements of environmental impact, infrastructure needs, and employment. The game scenarios are designed to assess systems thinking. Often named on lists of twenty-first century skills, systems thinking is also a cross-cutting concept in the Next Generation Science Standards (NGSS; NGSS Lead States, 2013). Essentially, it is the understanding of how various components of a system influence each other.

Table 14.1 Systems thinking learning progression from SimCityEDU

Level 1—Acausal
The player is not reasoning systematically about causes and effects
Level 2—Univariate
The player tends to focus on a single causal relationship in the system
Level 3a—Early multivariate
The player has considered multiple effects resulting from a single cause
Level 3b—Multivariate
The player has considered multiple causes in relation to their multiple effects
Level 4—Emergent patterns
The player attends to and intervenes on emergent patterns of causality that arise over time

Starting with a strong research-based theory or cognitive model is preferable (but not required) in the development of GBAs because it can provide clear hypotheses to design, categorize, and evaluate evidence that can be further explored through EDA. The aim is to jumpstart the design of GBAs using an initial psychological theory of students' likely changes in competency toward the learning goals during game play. This approach leverages existing models of learning and how peoples' understanding of concepts potentially progress through qualitative changes in a particular developmental sequence (e.g., learning progressions of how their thinking develops from simpler more univariate concepts to more complex interactive systems; c.f., Heritage, 2008). These learning progressions, or cognitive models, help to inform design and development of GBAs, but the models themselves are also subject to iterative refinement as data are collected during playtesting, mini-tryouts, pilots, and larger-scale studies. Following a review of existing conceptualizations of systems thinking, a learning progression for the construct was developed as part of the student model for the game. Table 14.1 presents a summary of the systems thinking learning progression used in SimCityEDU.

The examples described in this chapter relate to efforts to uncover evidence in players' game actions related to systems thinking. While SimCityEDU consists of four scenarios, discussion here focuses on the third, which requires players to balance maintaining enough power in the city with reducing air pollution. Players explore the city and find that coal plants are primary producers of pollution, while other industrial areas also contribute to the problem. Players can reduce pollution by bulldozing coal plants, but that will reduce power in the city. They can dezone industrial areas, but that alone will not result in large enough changes to please the city inhabitants (and get the player to a full three-star solution).

The process of analyzing the various actions players take in the game relies on the telemetry system of the game, or the remote collection of player actions and game states. Log files of telemetry data are collected for every game session and detail actions the player has taken in chronological order. The following sections seek to identify elements of game play that may provide insight into players' systems thinking using the principles of revelation, resistance, re-expression, and residuals with SimCityEDU data from 751 US middle school players who participated in beta testing of the game.

2 Revelation

Revelation refers to Tukey's (1977) statement that "The greatest value of a picture is when it forces us to notice what we never expected to see" (p. vi). Graphics are the primary tool for the exploratory data analyst. Graphical representations can display large amounts of information using relatively little space and expose relationships among pieces of information better than other representations. Here we are talking not about visualization for public display, but for finding patterns in relationships. Tools for this include things like boxplots and scatterplot matrices (a grid of scatterplots similar to a correlation matrix except with graphs) in addition to interactive graphics that allow the analyst to explore relationships with a few clicks. For example, a scatterplot may reveal a cluster of outliers. Interactive graphics allow the analyst to highlight them on the screen and examine their values on other variables to further understand what differentiates this group.

The initial goal of data analysis should be to become very familiar with the data. Instead of beginning an analysis by producing tables of descriptive statistics, followed by a big correlation matrix, EDA suggests beginning by looking at histograms, followed by scatterplots, scatterplot matrices, and boxplots. Let's take an example from the third scenario of SimCityEDU. The successful player will find out that coal power plants are the biggest pollution generators as well as the major energy producers and, therefore, both important and destructive for the city. Further, students engaged with this scenario need to discover that there are other energy sources available for them, such as solar or wind plants that are environmentally friendly. They have to figure out how replacing of coal power plants with green energy sources will allow them to reduce pollution while maintaining power in the city. We began with a rough hypothesis that just bulldozing coal plants without placing green energy would indicate a lower level of systems thinking because it indicated players were only considering a single effect of coal plants (namely pollution) rather than the multiple effects (power and pollution).

One of the first types of analyses is simply to examine the different actions and outcomes of game play. A common next step is to run the means and standard deviations, resulting in a table like the one in Table 14.2. Pollution is the final amount of

Table 14.2 Means and standard deviations of select outcomes and actions from SimCityEDU

Outcomes and actions	Mean	SD
Pollution	15,956,941	18,258,294
Bulldoze coal	3.17	1.952
Place new coal	0.20	0.745
Turned off coal	0.41	0.970
Turned on coal	0.17	0.678
Place wind/solar	2.58	2.366
Bulldoze wind/solar	0.26	1.008
Turned on wind/solar	0.07	0.370
Turned off wind/solar	0.09	0.430

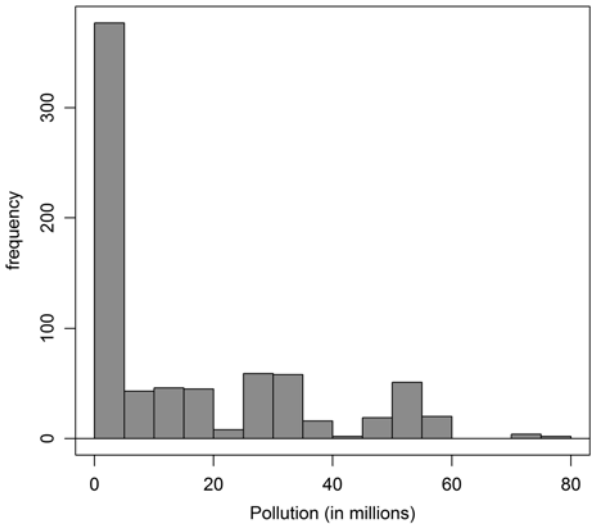


Fig. 14.1 Histogram of final pollution values

pollution in the city. Bulldozing refers to how players can eliminate buildings in the city (they use a bulldozing tool to knock them down). Placing refers to putting a new building in the city. Turning off and on are options to allow the energy plants to be active or not. Coal refers to coal plants while solar/wind refer to the alternative energy power plants available. So, on average, players knocked down 3.17 coal plants during their play, for example.

This representation does not tell us about the distribution or the outliers. However, a histogram like that in Fig. 14.1 for pollution does a better job showing these. If researchers start with visualizations first, they will better be able to interpret what numbers like those in Table 14.2 are indicating (or not indicating).

Here we see that pollution is quite skewed towards low values and actually appears to be trimodal. These three apparent groups in the outcome variable were not initially expected. The game was designed such that lower levels of pollution should be indicative higher levels of systems thinking, as players need to understand the system in order to successfully lower pollution without driving the city into a power failure. The identification of three groupings of pollution scores, however, was not intentional and raises questions about how game actions relate to these outcomes, and to systems thinking. While the groupings do not mean that the intended relationship of lower pollution to higher levels of systems thinking do not hold, it does mean that we must determine whether these groupings are artifacts of game design or whether they map to the levels of systems thinking. The latter would be a beneficial, but unexpected, result.

In Fig. 14.2, we can see the distributions of some of the other game actions. Note that bulldozing 4–6 coal plants is common. There are only six possible coal plant/generators in the original city, so anyone who bulldozed more than that must have placed new ones down. Understanding both the skew of the distributions and the location of outliers will lead into the re-expression and resistance work to follow.

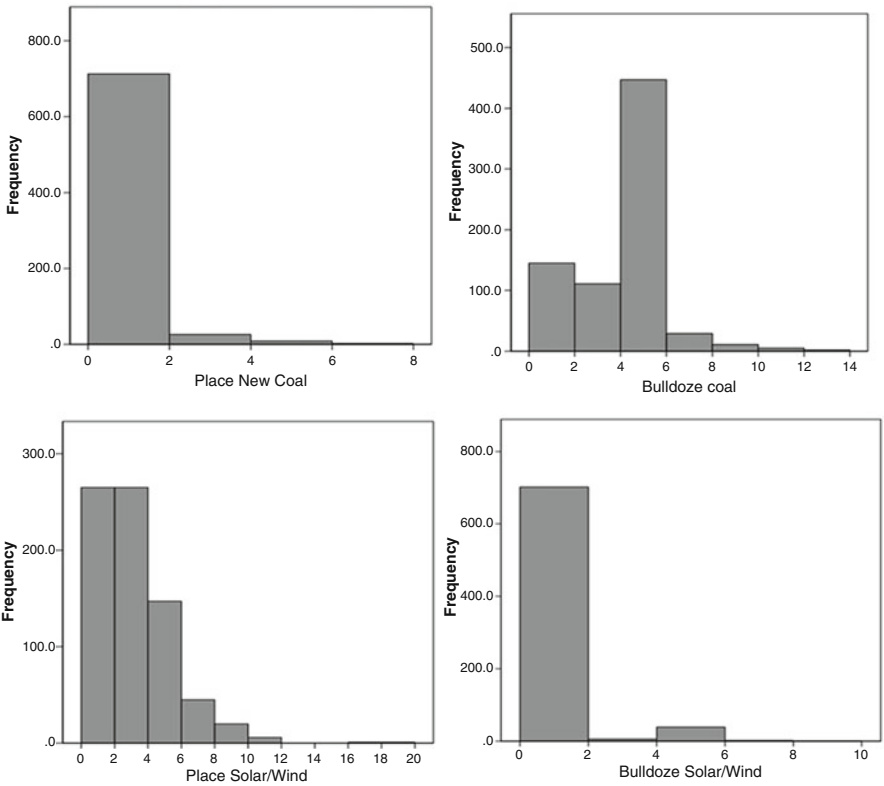


Fig. 14.2 Histograms of placing and bulldozing energy sources

Table 14.3 Correlation matrix among coal events, alternative energy events, and end state pollution

	Pollution	Bulldoze coal	Place new coal	Turned off coal	Turned on coal
Pollution	1.00				
Bulldoze coal	−.54	1.00			
Place new coal	.07	.35	1.00		
Turned off coal	−.03	−.31	−.04	1.00	
Turned on coal	.06	−.19	−.01	.82	1.00

Once we looked at this univariate information, we started looking at relationships between variables. A common technique to examine bivariate relationships is the creation of a correlation matrix like that in Table 14.3.

This suggests a moderate negative correlation between pollution and bulldozing coal, but not with other variables related to coal levels. However, the numbers themselves do not provide information about the patterns of relationship (for example, linearity and nonlinearity). To see those, scatterplot matrices such as the ones in Fig. 14.3 are helpful.

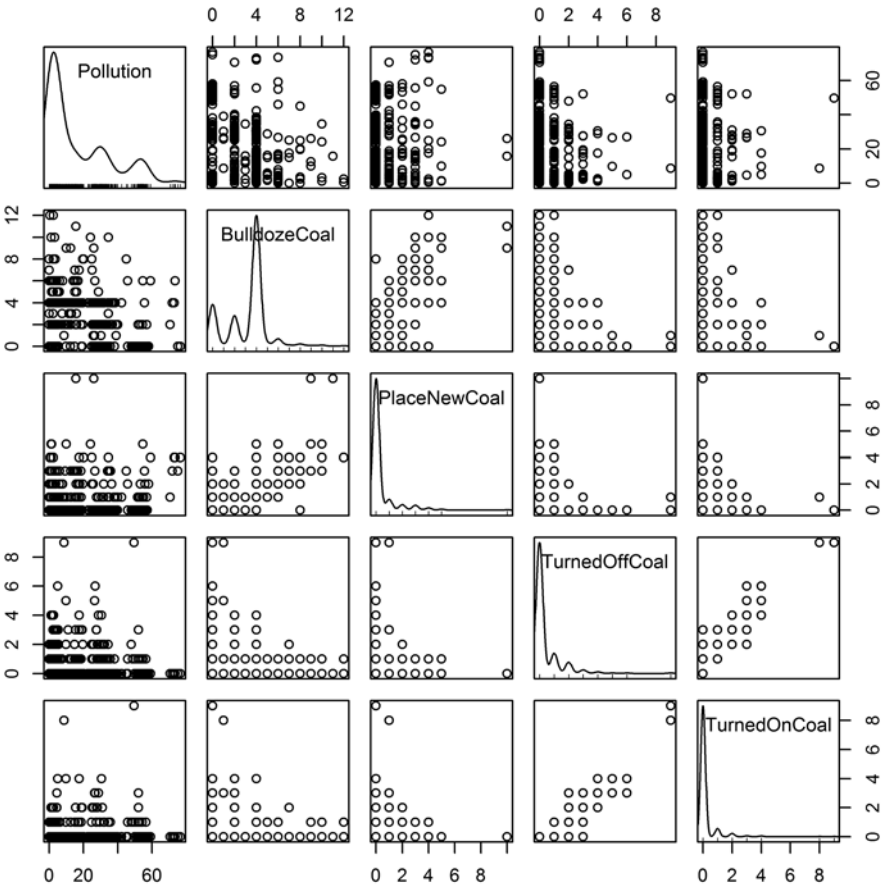


Fig. 14.3 Scatterplot matrix of relationship between coal activities and pollution

Figure 14.3 shows all of the actions that can increase or decrease the amount of coal production in the city. This matrix works like a correlation matrix such that each square is the scatterplot of the row and column variable with distributions of each variable on the diagonal. So the second box on the top row shows us the relationship between bulldozing coal and pollution. One thing that is apparent looking at this box is that there are some players that do not bulldoze any coal plants, but still end up with low pollution. These will require more investigation. Looking at the far right column, the fourth box down shows the relationship between turning coal plants off and on. When a player enters the game, all of the coal plants are on. This graph suggests that many of the players that turn a plant off proceed to turn it back on again. This behavior coincides with observations made during play testing that the turning off behavior is often a "testing" behavior in which the player can test the effect of turning a coal plant off without the permanency of bulldozing it. However, it is clear that this action is often reversed by turning it back on. Therefore, when we are looking

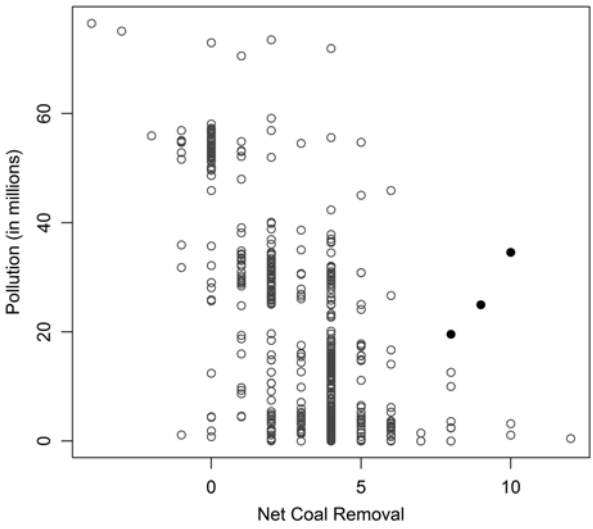


Fig. 14.4 First attempt to visually analyze relationship between net coal removal and pollution

to determine the total amount of coal removal, what we actually need is a measure of net removal that takes into account the reversal of removal actions.

We therefore created a variable adding the number of bulldozing coal and turning off coal actions and then subtracting the placing new coal and turning on coal actions to get a measure of net coal removal. The first time we created this variable and plotted the net coal removal against pollution, the result was that shown in Fig. 14.4.

In analyzing this, we were drawn to the three filled in black data points in the lower right of the figure. These were apparently individuals who had high net coal removal but continued to have relatively high pollution values. In order to search for other explanations for their pollution values, we returned to their log files. Rather than finding some other variable to explain the high pollution, we found that they had in fact placed additional coal plants that had not been properly coded in the automated scripts that clean the data. Here our visualizations helped us identify an error in our own data cleaning processes, and avoiding being fooled by incorrect data. Going back and fixing the coding of these values yielded the graph in Fig. 14.5.

3 Resistance

Because a primary goal of EDA is to avoid being fooled, resistance is an important aspect of using EDA tools. Resistant methods are methods that are less sensitive to large disruptions in small parts of the data (Mallows, 1983). Thus, they help us reduce the effects of extreme or unusual data. Note that this is different than robustness in that robustness deals with the ability of a statistic to give adequate estimates when assumptions are violated. Resistant methods are those that generally do not have these assumptions. In general, there are three primary strategies for improving

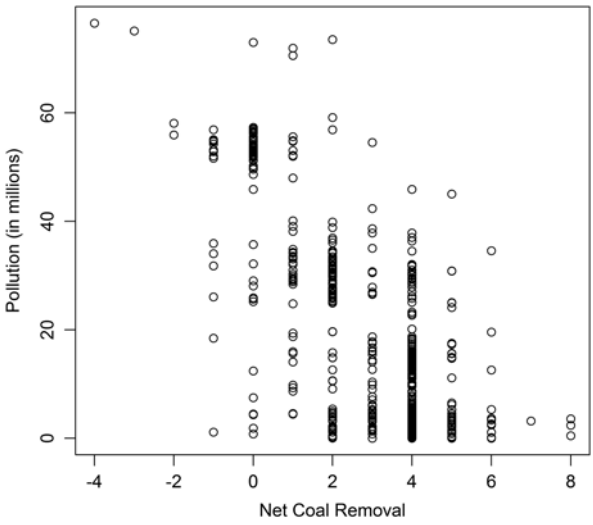


Fig. 14.5 Corrected scatterplot of relationship between net coal removal and pollution

resistance. The first is to use rank-based measures (e.g., the median) and absolute values, rather than measures based on sums (e.g., the mean) or sums-of-squares (such as the variance). While the mean has a smaller standard error than the median, and so may be an appropriate estimator for many confirmatory tests, the median is less affected by extreme scores or other types of perturbations that may be unexpected or unknown in the exploratory stages of research. For measures of spread, the interquartile range is the most common resistant method. The second general resistance building strategy is to use a procedure that emphasizes more centrally located scores, and uses less weight for more extreme values. This category includes trimmed statistics in which values past a certain point are weighted to zero, and thereby dropped from any estimation procedures. A third approach is to reduce the scope of the data one chooses to model on the basis of knowledge about extreme scores and the processes they represent. Depending on the application and the intended use of results, different methods will be appropriate in different situations.

3.1 Dealing with Outliers

Because an important goal of EDA is to develop understandings and descriptions of data, it is important to recognize that the data arise in specific contexts and contain background assumptions, even when these assumptions are unrecognized. This context and background can help us determine how to deal with outliers. Do we keep them or pull them out?

The fundamental question to ask is: Do we know something about these observations that suggests they come from a different process than the process we are seeking to understand? In games, numerous unintended processes could lead to outlying

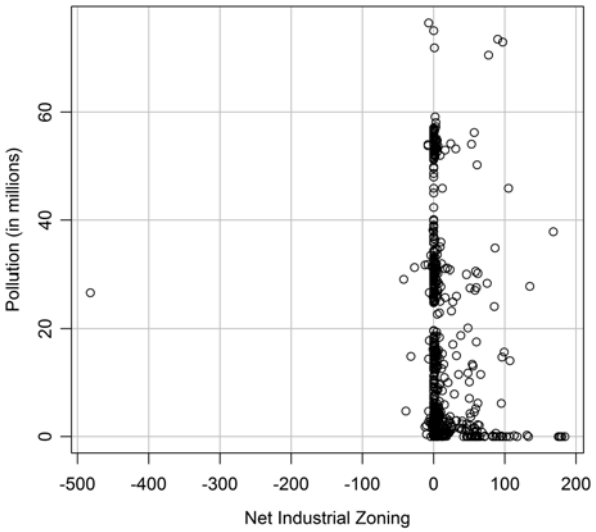


Fig. 14.6 Scatterplot of pollution and net industrial zoning

values: failure to understand instructions, exploring the environment, following their own goals, failure to pay attention to the task, or equipment or data failures. Games often encourage exploration and player agency, which means that players can often be observed doing things unrelated to the processes we wish to observe.

As an example, when we create a scatterplot of the net industrial zoning (another factor that should reduce pollution) versus pollution, we get the plot in Fig. 14.6. There is clearly one outlier who removed more than 500 industrial zones from the city. Further examination of this individual's log file revealed this individual also bulldozed 349 residential structures (median for the sample = 4) and 64 commercial structures (median = 3) while also dezoning 556 residential areas (median = 3) and 171 commercial areas (median = 0). This is an individual who appears to be seeking to destroy or eliminate most of the pre-built city. This is clearly a different goal than that intended and means we really cannot make any inferences about this individual's level of systems thinking. As a result, this is a case where it is justifiable to remove an outlier.

Alternately, when we look back at the scatterplot in Fig. 14.5, we could call the two values in the upper left outliers. They have higher pollution than any other players and lower net coal removal. However, these players' frequencies on other bulldozing and zoning variables are consistent with other players. There is no evidence that these players are not attempting to reduce pollution, they just are not doing it very well. Therefore, they were left in the sample, but the inclusion of their more extreme values point to the need for reporting of medians and interquartile ranges when reporting descriptive statistics. The most important aspect in either case is that a careful and detailed description of the full data, the reduced data, and the impact of the outlying data be reported. Unfortunately, the extremely terse descriptions seen in a lot of research reporting is inconsistent with this highly descriptive approach.

4 Re-expression

Data often come to the exploratory data analyst in messy, nonstandard, or simply not-useful ways. This may be overlooked if one assumes the data distributions are always well behaved, or that statistical techniques are sufficiently robust that we can ignore any deviations that might arise, and therefore skip detailed examination. In fact, it is quite often the case that insufficient attention has been paid to scaling issues either in advance, or during the modeling phase, and it is not until the failure of confirmatory methods that a careful examination of scaling is undertaken. Addressing appropriate scaling in advance of modeling is called re-expression and is a fundamental activity of EDA. Recently, advances in modeling have resulted in the ability to model distributions and nonlinearity, but still require careful consideration of underlying distributions in order to specify the appropriate model. Re-expression here refers solely to attempts to address the scaling of the data, as opposed to smoothing, for example, which aims at reducing the variability of the data.

The distribution most commonly "assumed" by statistical tests is the "normal" distribution. In EDA, the term "normal distribution" is avoided in favor of "Gaussian distribution" to avoid the connotation of prototypicality or social desirability. A Gaussian shape is sought because this will generally move the data toward more equal-interval measurement through symmetry, will often stabilize variance, and can quite often yield forms of the data that lend themselves to other modeling approaches (Behrens, 1997).

4.1 Re-expression Prior to Modeling

Although mathematically equivalent to what is called transformation in other traditions, re-expression is so named to reflect the idea that the numerical changes are aimed at appropriate distributions rather than radical change. An appropriate re-expression can often be found by moving up or down the ladder of re-expression (Tukey, 1977). The ladder of re-expression is a series of exponents one may apply to original data that show considerable skew. Recognizing the raw data exists in the form of X1, moving up the ladder would consist of raising the data to X2 or X3. Moving down the ladder suggests changing the data to the scale of X1/2, −X-1/2, −X-1, −X-2, and so on. The position on the ladder occupied by X0 is generally replaced with the re-expression of log(X), where the log is usually either taken to be the base 10 logarithm or the natural logarithm; the choice between them is arbitrary but may be made for interpretation. Gelman and Hill (2007) for example, suggest that the base 10 logarithm yields easier interpretation of data while the natural logarithm yields easier interpretation of coefficients in models. Note that the Box-Cox power transformation is one more formal method by which to search for and apply the best means of re-expression.

To choose an appropriate transformation, one moves up or down the ladder (i.e., takes each data point and applies the appropriate exponent) toward the bulk

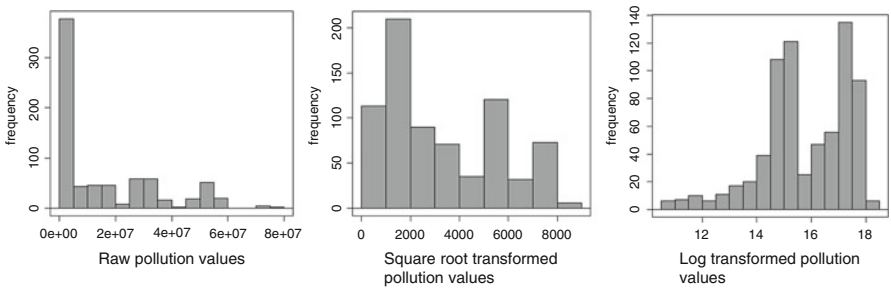


Fig. 14.7 Re-expression of end state pollution variables

of the data. This means moving down the ladder for distributions with positive skew and up the ladder for distributions with negative skew. To demonstrate this process, we can examine the distribution of the end state pollution values. These are initially highly skewed and were re-expressed with both a square root and log transformations (Fig. 14.7). The square root transformation shifted the distribution somewhat to the right but still leaves some skew (skew: 0.48). However, the log of pollution shifted the data too far, resulting in a negatively skewed distribution (skew: −2.91).

A common objection to re-expression is that the results of analyses involving re-expressed variables are difficult to interpret. This is true in some cases, however, we wish to provide an example of interpretation of log re-expression in regression to demonstrate that this should not be a barrier for some of our most common analyses. In the situation where the dependent variable is re-expressed as a log of the original variable while the independent variables are not, we say that a one unit change in the independent variable yields a 100*coefficient percent change in the dependent variable. In the case where the independent variable is re-expressed as a log but the dependent variable is unchanged, we interpret the result as a 1 % change in the independent variable results in a coefficient/100 change in the dependent variable. When both the independent and dependent variables are re-expressed as logs, we can interpret the regression result to mean that a 1 % increase in the independent variable leads to a coefficient percent increase in the dependent variable. It should be noted that re-expression alters the relative distance between data points. So, although the points all remain in the same order, there is a loss of information that may be undesirable when those distances are meant to be interpretable, such as might be the case with variables such as age or GPA (Osborne, 2002).

Although some researchers may reject the notion of re-expression as "tinkering" with the data, our experience has been that this view is primarily a result of lack of experience with the new scales. In fact, in many instances individuals use scale re-expressions with little thought. For example, the familiar practice of using a proportion is seldom questioned, nor is the more common re-expression to z-scores. Many common measurements, such as the Richter scale and decibel are transformations.

4.2 Modeling Distributions

The re-expression discussed up to this point has involved re-expression of individual variables prior to model fitting. This work is important in that it builds familiarity with the data, helps to understand different possible strategies, and suggests possible approaches for picking a computational model for an analysis. Rodgers (2010) discusses a "quiet methodological revolution" (p. 1) in which the traditional null hypothesis–testing paradigm is replaced with one of building, evaluating, and comparing models. The focus of the new paradigm is on developing models that best fit the data, rather than manipulating the data to fit the assumptions of a test of a null hypothesis and may involve re-expression to better bring out relationships.

After completing the scale-motivated methods discussed earlier, exploratory analyses often take advantage of the strengths of generalized linear models. For example, while a binary variable may be transformed to a series of logits for early data exploration, the development of a predictive model is most likely to be accomplished using a logistic regression form with all the availability of predictive values, residuals, and so forth available in common generalized linear models. In other words, data may be re-expressed for some analyses, but also left in its raw form and models incorporating the non-Gaussian distributions used. For example, count data are commonly analyzed using Poisson (log-linear) models without initial re-expression of the data. Weighted least squares can be used when variability is not constant across groups (heteroscedasticity). Gelman and Hill (2007) provide excellent examples on the application of generalized linear models following approaches largely or altogether consistent with the views expressed here. Finally, nonparametric methods can be explored, although often at the expense of power and loss of information from interval level scales.

5 Residuals

George Box (1976) succinctly summarized the importance of aligning model choice with the purpose of the analysis writing: "All models are wrong, some are useful" (p. 3). Residuals allow us to understand how our models are wrong. This emphasis on residuals leads to an emphasis on an iterative process of model building: A tentative model is tried based on a best guess (or cursory summary statistics), residuals are examined, the model is modified, and residuals are reexamined over again.

It is worth a pause here to describe how these models are built. In a traditional research view, models are developed from the hypotheses of experts with domain knowledge and the existing research base. However, in GBAs we often have very weak or nonexistent hypotheses about the relationships among variables of interest. As a result, it is prudent to examine recent advances in methods of model building. For example, researchers can submit data to Kaggle and set up a competition among data scientists to find the best models of the data, essentially crowdsourcing

model building. Alternately, statistical techniques such as symbolic regression can be used to discover the relationships among variables in a model. In using any of these traditional or new techniques, a key is understanding not just the fit of the model, but where misfit is occurring.

In statistics, we use the term residual to mean what is left unexplained by the predictor(s). If you are trying to predict someone's test score by how much they studied, you are going to be wrong, for some people by a little and for some people by a lot. That amount you are off is what is "left over" of the test score after the effect of study time is accounted for. It is the residual. Different models will lead to different patterns of residuals. It is not just a case that some are big and some are small, but that when they are graphed, we can see patterns. In many models, there are assumptions about residuals. For example, in linear regression, a well-fit model will have residuals with a mean of 0 and variance should be constant. However, even without specific assumptions, examining the pattern of error terms, can yield information about how models fit the collected data.

In the EDA tradition, residual is not simply a mathematical definition, but a foundational philosophy about the nature of data analysis. The primary focus of EDA is on the development of compact descriptions of the world. However, these descriptions will never be perfect so there will always be some misfit between our model and the data, which really means a misfit between our model and the world.

In the third scenario of SimCityEDU, we wanted to examine the factors that led to decreased pollution outcomes in the game (hypothesizing that players who ended up with lower pollution while maintaining power had a better understanding of the system). In order to test the variables explored above, we ran a linear regression model predicting the square root of pollution from the net coal removal, net industry removal, and net alternative energy placement. The model was significant, $F(3, 745) = 303.5$, $p < .001$, $R^2 = .55$, Cohen's $f^2 = 1.22$. Importantly, all three predictors were significant, indicating they all contribute to pollution values above and beyond the other predictors, and engaging in those activities is likely related to understanding of the system. These three variables plus the end state are the beginnings of evidence we will include in our measurement model.

While the model is statistically significant, we should not stop there. We can graph the predicted pollution outcomes for each person versus the residuals (see Fig. 14.8).

Looking at the graph, we see that there is a clear pattern in which lower predicted values of pollution have higher residuals and higher values of pollution have smaller residuals. A biased homoscedastic pattern such as this suggests there is likely an unmodeled predictor variable.

Based on this information, we will want to adjust our model. We can do this in a number of ways. We might try to statistically model the pattern. In this case, we tried a general linear model using raw pollution values and a Poisson distribution. This yielded an even more extreme linear pattern. Our next path will be to find another predictor to add to the equation. It may be that the group that is under-predicted did something else to decrease pollution in the city. Going back to exploratory mode and/or using some other data mining techniques might uncover this.

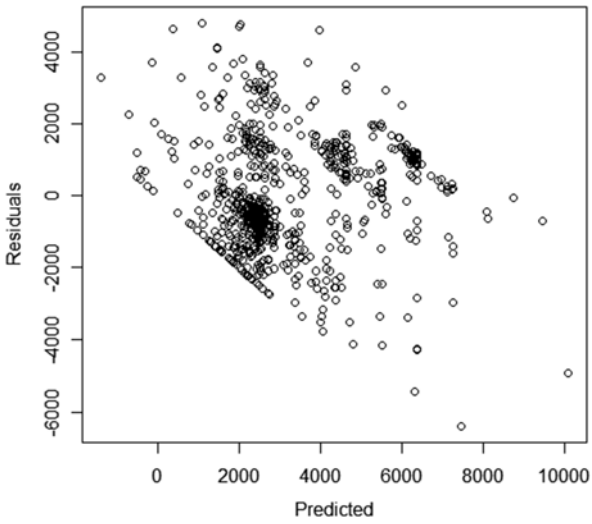


Fig. 14.8 Scatterplot of actual pollution versus residuals

There are two cautions with this process. First, there is a point of diminishing returns where the improvements made to the model no longer have meaningful impact on the decisions to be made. For example, the slightly greater precision in estimation of ability may not be useful in informing instructional decisions. Second, going down the iterative exploratory road fits a model to a particular data set, and confirmation on independent data would be required.

6 Psychometric Techniques

To finish the discussion of evidence models in GBAs, we will briefly review how the pieces of evidence identified in EDA are combined using a measurement model in order to estimate players' levels of systems thinking. The EDA processes we saw above may yield everything from action counts to times to final scores as evidence fragments. While these individually may be interesting, we must also find a way to combine these disparate pieces of information to estimate the values of the latent traits we are ultimately interested in assessing. This is the work of the measurement model.

The simplest psychometric models are classical test theory (CTT) models or observed score models, in which scores based on observable variables are added. CTT works well when the multiple measures at issue are similar pieces of evidence about the same thing—in familiar assessments, for example, correctness across many similar test items; in GBAs, this would correspond to independent attempts at similar problems, as long as learning is negligible across those attempts. With familiar tests, CTT models also prove serviceable for collections of unlike items—as long as the collection doesn't change. Since CTT addresses the overall score, changing game scenarios or player actions changes the meaning of the scores; it

does not lend itself to the rapid versioning of games or their mix-and-match character. In general, CTT does not work as well for situations that are more complicated in any of several ways: for example, where the evidence comes in different forms, has dependencies among some of its pieces, pieces depend on different mixes of skills in different combinations, proficiencies are changing across the course of observation, or different players contribute different amounts or different types of evidence. In the example above, we have information such as net coal removal events and total end state pollution. It would not make sense to simply add these values up. Latent variable models were invented to deal with assessments with these features.

Commonly used latent variable models used in educational measurement include item response theory (IRT; Yen & Fitzpatrick, 2006) and diagnostic classification models (von Davier, 2005). More detail about latent variable models can be found in Mislevy et al. (2014). Developed in traditional assessment environments, these models often have constraints on independence of observations and single dimensionality of observations that are routinely violated in GBA. While modifications of the models, such as multidimensional IRT have been developed, the multidimensional, dependent evidence with polytomous or continuous observations continue to challenge these items. Bayesian inference networks offer another option and have shown to be useful in complex assessment systems with nontraditional evidence (Almond & Mislevy, 1999; Mislevy & Gitomer, 1996; VanLehn, 2008). There is no need to pick "a" model from among them to use in GBA, because different kinds of observable variables (counts, strategy usage, features of an system diagram) can all be modeled as depending on the same latent variables by using appropriate conditional probability distributions (link functions). Furthermore, it is sometimes useful to have multiple models running in parallel, or to have them running at different levels of the hierarchical organization of GBA interactions.

We focus here on Bayesian inference networks and provide a numerical example to give some insight into how the model works in GBA. Bayesian inference networks, or Bayes nets for short, are a broad class of models for interrelationships among categorical variables. They can express or approximate the various latent-variable models mentioned above, and are particularly well suited to flexible combination of modules that express recurring relationships among kinds of evidence or between evidence and proficiencies (a characteristic that serves well in domains such as jurisprudence, intelligence analysis, and medical diagnosis; Schum, 1994). The model enables us to take advantage jointly of information from theories about a learning domain, from design strategies, and accumulating data from players. At the beginning, we posit models that reflect our initial beliefs about the targeted aspects of proficiency and the features of situations (tasks) that will evoke them. We build these hypotheses into the forms and the parameterizations of the models. By modeling conditional probabilities in terms of parameters, we can express our initial expectations as prior probability distributions for the parameters. As data arrive, Bayesian machinery allows us to get increasingly improved estimates of the model parameters and to examine where and how well the data fit the model. This information helps us fine-tune models to better manage evidence, or to modify

game situations to provide better evidence. This is a particular advantage of Bayesian networks; as long as the student model variables (SMVs) remain the same, it is straightforward to incorporate additional forms of evidence, such as new evidence fragments discovered from educational data mining (EDM) or new game levels added to the game.

Koenig, Lee, Iseli, and Wainess (2010) and Shute (2011) illustrate the use of Bayes nets in GBA, with ECD as the design framework. VanLehn (2008) provides a good overview for related uses in intelligent tutoring systems. An example from the Sierra Madre challenge in SimCityEDU illustrates key ideas.

6.1 A Numerical Example

Figure 14.9 gives a numerical example of a part of a Bayes net for the third scenario. As we will see, Bayes nets generally require categorical states for the observable variables. Recall that in the above analysis the final pollution state appeared to have a trimodal distribution. Three groups were identified in the pollution result and combined with final power state to yield five levels of an End State observable variable. Similarly the net coal removal variable and net industry zoning variables identified above were combined with the net alternate energy placement variable to form a Remove Replace variable. Systems Thinking is the latent SMV and Remove Replace and End State are two observable variables, as shown in Fig. 14.9 (this is a small piece of the Bayes net for demonstration purposes). Recall that Systems Thinking has five levels. However, the design of the game targeted gathering evidence for the first four. Therefore, the top two levels are collapsed given that with the evidence

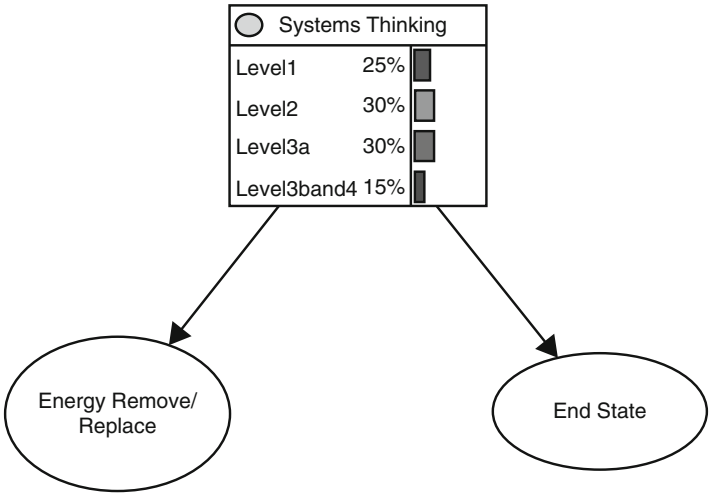
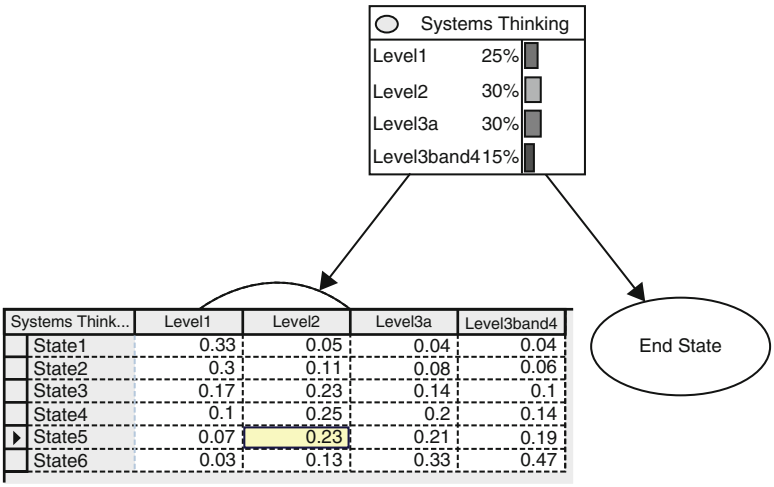


Fig. 14.9 Bayesian Network with no observed evidence



Systems Think...	Level1	Level2	Level3a	Level3band4
State1	0.33	0.05	0.04	0.04
State2	0.3	0.11	0.08	0.06
State3	0.17	0.23	0.14	0.1
State4	0.1	0.25	0.2	0.14
State5	0.07	0.23	0.21	0.19
State6	0.03	0.13	0.33	0.47

Fig. 14.10 Probability table linking evidence node to student model variable

available in the game we cannot differentiate between the two. Figure 14.9 shows the prior probabilities we assign to a student being at these levels, before observing her performance. Note that although it is traditional for diagrams to display latent variables as circles and observed variables as squares, Bayesian Network software consistently displays all variables as circles and uses squares when displaying the probability distributions, as seen in Fig. 14.9. The values shown there represent beliefs that correspond to how we expect the game to be used. That is, most players would be at level 1, 2, or 3a with respect to this context and content, and not at 3b or 4; however, without evidence, there are near equal probabilities that a player is at level 1, 2, or 3a.

We then create probability tables, like that shown in Fig. 14.10, that list the probability of observing each category or state of energy remove and replace given a level of systems thinking. So, for example, someone at level 1 of the systems thinking progression would have a .33 probability of not removing any coal or industry (State 1). The numbers for prior probability and conditional probabilities were first justified in terms of what we know about the situation—expectations based on knowing the kinds of students who would be players, research on Systems Thinking, and the numbers in the example are initial expert-opinion refined by data from a small scale try-out test. As the general release of the game brings in much large volume of data, the Bayes net allows for coherent updating of the conditional probabilities (Mislevy, Almond, Yan, & Steinberg, 1999). The model also allows for comparing the patterns in the data with the patterns the model can express, so that the model or the data-gathering situations can be improved (Levy, 2006; Williamson, Mislevy, & Almond, 2000).

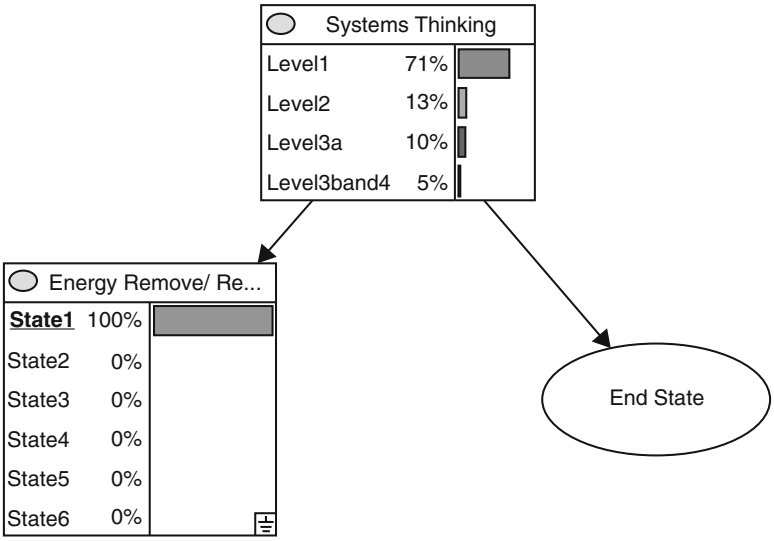


Fig. 14.11 Bayes Net after observing play

We created a similar probability table for the variable EndState. Once the probability tables are created, we can use the Bayes nets to estimate probabilities, as shown in Fig. 14.11. For example, if we see that someone has not removed any coal plants or rezoned any industrial areas, we can then update their probabilities of being at each level of the systems thinking progression. In this case, the updating results in the estimate that there is a .76 probability that the player is at level 1 (Acausal Thinking) in the progression. In this way, we are able to link in-game actions to estimates of levels of the learning progression.

7 Conclusions

The goal of the exploratory analysis of SimCityEDU was to identify potential pieces of evidence in game play related to systems thinking. The tools presented here were useful in identifying these "fragments" of evidence that could then be combined via statistical tools such as Bayesian networks. They allowed us to identify errors in our data process, suggested how actions might be used by players (e.g., turning off coal plants as a test), identify outliers and assess their inclusion in models, and judge whether our efforts to identify meaningful variables was complete. The methods of EDA are summarized in Table 14.4.

Table 14.4 Summary of 4R's of EDA (based on Hoaglin et al., 1983)

	Definition	Example
Revelation	Uncovering the unexpected, most often through visualization	Identification of a cluster of players whose actions led to unexpected game outcomes
Resistance	Using methods that are not overly influenced by extreme or unusual data	Identification of players whose actions are so different than average that they are likely pursuing a different goal in game play, suggesting we should not make inferences about their skill based on our known evidence rules
Re-expression	Ensuring match between data distributions and modeling techniques	Use of square root or log-transformed variables to better fit models
Residuals	Evaluation of where models do not fit the data, encouraging iteration	Identification of overprediction of a model at the lower end of a scale, suggesting variables are missing from the model

We believe that EDA offers a complementary approach to other analysis traditions. For example, EDM is a group of methods aimed discovering novel and useful information from large amounts of educational data. Baker and Yacef (2009) identified the following five areas of work characteristic of EDM: prediction, clustering, relationship mining, distillation of data for human judgment, and discovery with models. It is our position that EDA allows for intimacy with data prior to the use of these more complex methods. In our experience, practitioners of EDM often wait until their models fit poorly to begin investigating the issues of data familiarity and distribution discussed here. In addition, EDA serves as a theory-generating process which can inform the data mining models being built (for example, informing the list of features used in building automated detectors). We believe beginning with EDA techniques would likely result in better fitting EDM models and more thorough understanding of results. The use of residual techniques will lead to better evaluation of the resulting models as well. In the SimCityEDU project, analysis will likely move to EDM techniques in an attempt to identify other variables involved in the prediction of pollution scores.

The work of identifying evidence from GBAs ultimately requires cycles of exploration, hypothesis generation, and confirmation. While EDA is likely good practice in analysis of all kinds of data, the generally weak initial hypotheses about links between game play and evidence combined with the open nature of game play in GBAs make GBA data a prime candidate for the use of the techniques. Our psychometric techniques have progressed to the extent that we can receive the traditional correct/incorrect data, fit our established models, and review the output with known methods to examine fit. However, GBAs result in new work products and new kinds of evidence that do not easily translate into these techniques. By looking back to Tukey's Exploratory Data Analysis tools, we find a framework and powerful tools to lead us forward in analyzing our new game-based assessment data.

References

Almond, R. G., & Mislevy, R. J. (1999). Graphical models and computerized adaptive testing. *Applied Psychological Measurement, 23*, 223–237.

Baker, R. S. J. D., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. *Journal of Educational Data Mining, 1*, 3–16.

Behrens, J. T. (1997). Principles and procedures of exploratory data analysis. *Psychological Methods, 2*(2), 131–160.

Behrens, J. T., DiCerbo, K. E., Yel, N., & Levy, R. (2012). Exploratory data analysis. In I. B. Weiner, J. A. Schinka, & W. F. Velicer (Eds.), *Handbook of psychology: Research methods in psychology* (2nd ed., pp. 34–70). New York: Wiley.

Box, G. E. (1976). Science and statistics. *Journal of the American Statistical Association, 71*, 791–799.

Gee, J. P. (2003). *What video games have to teach us about learning and literacy*. New York: Palgrave Macmillan.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/ hierarchical models*. New York: Cambridge University Press.

Heritage, M. (2008). *Learning progressions: Supporting instruction and formative assessment*. Washington, DC: Council of Chief State School Officers.

Hoaglin, D. C., Mosteller, F., & Tukey, J. W. (1983). *Understanding robust and exploratory data analysis*. Hoboken, NJ: Wiley.

Klopfer, E., Osterweil, S., & Salen, K. (2009). *Moving learning games forward: Obstacles, opportunities, and openness*. Cambridge, MA: The Education Arcade.

Koenig, A. D., Lee, J. J., Iseli, M., & Wainess, R. (2010). *A conceptual framework for assessing performance in games and simulation.* (CRESST Report 771). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Levy, R. (2006). *Posterior predictive model checking for multidimensionality in item response theory and Bayesian networks*. Doctoral dissertation, University of Maryland at College Park.

Mallows, C. L. (1983). Data description. In G. E. P. Box, T. Leonard, & C.-F. Wu (Eds.), *Scientific inference, data analysis, and robustness* (pp. 135–151). New York: Academic Press.

Mislevy, R. J., Almond, R. G., Yan, D., & Steinberg, L. S. (1999). Bayes nets in educational assessment: Where the numbers come from. In *Proceedings of the fifteenth conference on uncertainty in artificial intelligence* (pp. 437–446). San Francisco: Morgan Kaufmann.

Mislevy, R. J., Behrens, J. T., DiCerbo, K. E., Frezzo, D. C., & West, P. (2012). Three things game designers need to know about assessment: Evidence-centered design for game-based assessment. In D. Ifenthaler, D. Eseryel, & X. Ge (Eds.), *Assessment in game-based learning: Foundations, innovations, and perspectives* (pp. 59–84). New York: Springer.

Mislevy, R. J., Corrigan, S., Oranje, A., Dicerbo, K., John, M., Bauer, M. I., et al. (2014). *Psychometric considerations in game-based assessment*. New York: Institute of Play.

Mislevy, R. J., & Gitomer, D. H. (1996). The role of probability-based inference in an intelligent tutoring system. *User-Modeling and User-Adapted Interaction, 5*, 253–282.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2002). Design and analysis in task-based language assessment. *Language Testing, 19*(4), 477–496. doi:10.1191/0265532202lt241oa.

NGSS Lead States. (2013). *Next generation science standards: For states, by states*. Washington, DC: The National Academies Press.

Osborne, J. W. (2002). Notes on the use of data transformations. Practical Assessment, Research & Evaluation, 8 (6). Retrieved from http://pareonline.net/getvn.asp?v=8&n=6.

Rodgers, J. L. (2010). The epistemology of mathematical and statistical modeling: A quiet methodological revolution. *American Psychologist, 65*, 1–12.

Schum, D. (1994). *The evidential foundations of probabilistic reasoning*. New York: Wiley.

Shute, V. J. (2011). Stealth assessment in computer-based games to support learning. In S. Tobias & J. D. Fletcher (Eds.), *Computer games and instruction* (pp. 503–524). Charlotte, NC: Information Age.

Steinkuehler, C. A. (2004). Learning in massively multiplayer online games. In Y. B. Kafai, W. A. Sandoval, N. Enyedy, A. S. Nixon, & F. Herrera (Eds.), *Proceedings of the sixth international conference of the learning sciences* (pp. 521–528). Mahwah, NJ: Erlbaum.

Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.

Tukey, J. W. (1986). Data analysis, computation and mathematics. In L. V. Jones (Ed.), *The collected works of John W: Vol. IV. Philosophy and principles of data analysis: 1965–1986* (pp. 753–775). Pacific Grove, CA: Wadsworth. Original work published 1972.

VanLehn, K. (2008). Intelligent tutoring systems for continuous, embedded assessment. In C. Dwyer (Ed.), *The future of assessment: Shaping teaching and learning* (pp. 113–138). Mahwah, NJ: Erlbaum.

von Davier, M. (2005). *A class of models for cognitive diagnosis*. Research Report RR-05-17. Princeton, NJ: ETS.

Williamson, D., Mislevy, R. J., & Almond, R. G. (2000). Model criticism of Bayesian networks with latent variables. In C. Boutilier & M. Goldszmidt (Eds.), *Uncertainty in artificial intelligence* (pp. 634–643). San Francisco: Morgan Kaufmann.

Yen, W. M., & Fitzpatrick, A. R. (2006). Item response theory. In R. Brennan (Ed.), *Educational measurement* (3rd ed., pp. 111–153). Portsmouth, NH: Praeger/Greenwood.