CSSS 569: Visualizing Data and Models
Visualizing Event History Data

Rachel Castellano, Andrea Cancino Sáenz, Megan Erickson, Hitomi Kariya,
Morgan Wack, Nicolas Wittstock

24 February 2020

Event history analysis captures the longitudinal record of when events occur for a set of observations. In other words, this type of analysis describes, explains, or predicts the occurrence of events. In the biomedical field, practitioners commonly use event data to develop prognostic records for disease recurrence, mortality rates, and treatment outcomes. This is why event history analyses are commonly conflated with survival analyses, capturing the intuition that analysts are examining the occurrence of death or the rate of survival. Alternatively, in the political science discipline, event history data can capture political, social, or economic phenomena. It thus becomes clear that common biomedical assumptions of event history data cannot seamlessly map onto political science – variation in design and variable measurement could demand different types of event history methods. In using event history data, scholars must therefore make clear the underlying assumptions of the data, methods, visualization, and subsequent analysis.

As such, this memo proceeds as follows. In Section 1 we outline broad strategies for how scholars may communicate event history analyses in being explicit about event data terminology and creating clear and accessible visuals. In Section 2 we discuss the visual means of exploratory data analysis. Finally, in Section 3, we discuss how visualizations can aid our understanding of different modeling strategies.

1 Terminology and Accessibility

This memo will discuss how to properly understand and approach event history data in research. However, it is equally important to consider how to communicate this research in a form that is both appropriate and accessible. The hard science and social science disciplines both prominently use event history data, demonstrating that such analyses lend itself to a wide audience. On one hand, the prominence of this kind of data and methodology can open opportunities for cross-disciplinary input, ultimately lending itself to better research. On the other hand, discipline-specific jargon and implied assumptions can stymie these kinds of conversations. Good research relies on clear communication as much as it relies on carefully crafted models and analysis. This section will discuss strategies and tools that will make event history data analysis easier to understand both in written and visual formats.

1.1 Terminology

Scholars and analysts conduct event history analysis under a variety of names, including duration analysis, hazard modeling, and, in the sectors of health and biostatistics, survival analysis. For example, the public health literature may refer to "survival data" whereas engineering literature may cite "failure-time models." At first glance, this variation may seem unnecessary as they are all different names referring to the same thing – models that concern the timing and duration until the occurrence of an event Mills 2014). However, cluing in the reader as to what type of event is being researched can help clarify the research scope and assumptions.

While the term "survival analysis" is prominent in medical research, for example, this term may not serve the same utility if it was adopted in social science research. Survival implies that there is a binary of life and death – survival time is defined as the time starting from a predefined point (i.e. diagnosis of the disease) to the occurrence of the event of interest (i.e. death) – but what if the data are describing an event that can oscillate between these

two states? Data on treaty ratification may show patterns of countries reneging a treaty but then re-ratifying in the future; data on regime "survival" may show a similar pattern with some countries democratizing, then returning to authoritarianism, only to attempt democratization again in the future. Thus, "survival analysis" makes sense in the medical context, when researchers are dealing with literal instances of how long a patient will survive given a particular condition, but may complicate assumptions in the social science realm. Deciding which terms to use when describing event data methods should depend on the discipline and topic at hand, and what kind of literature it fits into.

Thus, best practices in event data analysis require scholars to explain terms and concepts as they relate to relevant fields of research and to avoid terminology that could be easily misinterpreted or misconstrued by the reader. Indeed, though the central terms and concepts of event history analysis are universal, they are often described in disparate terms across sectors. Of these terms, the following are required for comprehension of more complex models utilizing event history analyses:

Censoring. Censoring refers to a challenge created by incomplete data. As a result of the temporal structure of time history analyses it is common for events to occur outside the window of observation.

> *Right censoring.* If only the lower limit for the event is known, such as when subjects' birth dates are known, but they remain alive when a study ends are lost from follow-up, this is called right censoring. This is the most common form of censoring.
>
> *Left censoring.* If only the upper limit for the event is known, as would occur when someone entering a study is diagnosed with a disease, but the original date of exposure is unknown, this is referred to as left censoring (or left truncated). This type of censoring is uncommon, but still occurs across sectors utilizing event history analysis.
>
> *Interval censoring.* This relatively common type of censoring occurs when an event

occurs at an unknown point between two points (often observations). If someone were to test positive for an event between pre- and post-meetings, but the exact date of inoculation was unclear, this would be a case of interval censoring.

Hazard function. Often denoted by lambda (λ), hazard functions are essential components of event history analyses. Described elsewhere as the transition rate, instantaneous risk, or failure rate, the hazard function is the probability that an event occurs during a given interval of time, divided by the width of the interval. This can be denoted mathematically as follows where t is time:

$$h(t) = \lim_{\Delta t \to 0} \frac{Pr(t \leq T < t + \Delta t | T \geq t)}{\Delta t}$$

Survival function. A similar concept is the survival function, which as the name suggests signifies the probability that an event does not happen before a specified time t. This can be quickly denoted as:

$$S(t) = Pr(T \geq t)$$

Cumulative distribution function. The complement to the survival function is the cumulative distribution function, which reflects the probability that an event does occur before a specified time t. This can be denoted as:

$$F(t) = 1 - S(t) = Pr(T < t)$$

1.2 Accessible Visuals

As scholars supplement models with graphical aids, they must be mindful of how they construct their visuals and aim to achieve a high level of both information and clarity. For example, one must ensure that graphs have a title that is more descriptive than simply reading "Event History Graph," and axes labels must be easy to interpret. Below is an

example of a graph that shows the relative risk of death for migrants broken down by reason for migration Syse et al. 2018).

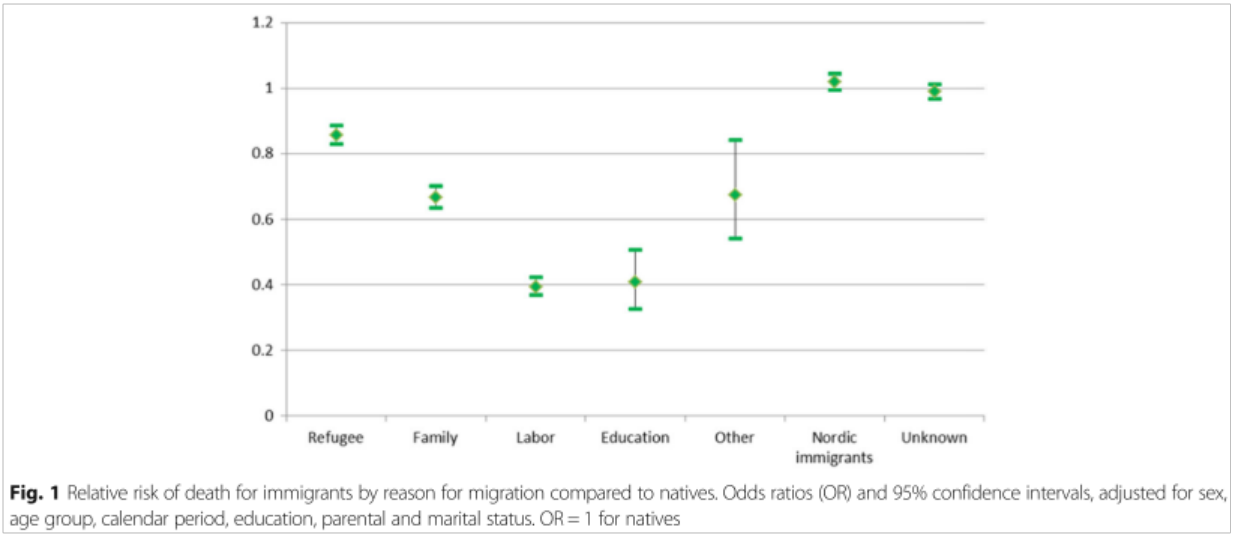


Fig. 1 Relative risk of death for immigrants by reason for migration compared to natives. Odds ratios (OR) and 95% confidence intervals, adjusted for sex, age group, calendar period, education, parental and marital status. OR = 1 for natives

Despite the descriptive caption of the figure, the graph itself is not intuitive to interpret. Neither axis is labeled. While the x-axis can be understood as the categories of reasons for migration, it is difficult to make sense of the scale used for the y-axis. It is not obvious whether this is relative risk in terms of percentages, or some different kind of scale.

Tools that allow for a more involved annotation of graphs include `R` packages like `ggplot2` or `tile`. `ggplot2` is user-friendly, but its declarative features may hinder readability by applying burdensome default themes and formatting. For example, the package gives a default gray background instead of a white background, creating unnecessary "noise" in the figure. Annotations in `ggplot2` may also be slightly more difficult to scale properly or manipulate the text in a way that maintains legibility. `tile`, on the other hand, is an imperative programming system, and thus it is slightly less intuitive than `ggplot2`. It allows users to control the exact construction of output, which is more work yet enables more design freedom avoids allows users to avoid sidestepping or modifying default settings.

2 Exploratory Analysis of Event History Data

Moving beyond ensuring the accessibility of event data terminology and visuals, uncovering patterns in event history analysis is an important step in exploring one's data. For example, one must consider censoring and missing data when determining what type of analysis and model is most appropriate for their data. One effective way to do this is through a heat map, such as the one below.

The primary purpose of a heat map in event history analysis is to better visualize when observations experience the event of interest. When does this occur? Is the event recurring? Are there patterns within different categories of the data set? Heat maps are also useful in uncovering trends around missing data. The heat map above shows whether or not countries report the criminalization of human trafficking survivors over a thirteen-year period. Before making this plot, it was clear that there was missing data, but the heat map uncovers an important pattern; missing data is much more likely before 2007. This plot also illuminates some regional patterns. For example countries in Northern and Western Africa, and Central and Southern Asia are more likely than not to criminalize survivors.

Below is another example of using a heat map plot to explore event history data. In this plot, Imai et al. (2020) introduce a treatment variation plot, which visualizes the variation of treatment across space and time. This plot is intended to help researchers build an intuition about how comparison of treated and control observation can be made.

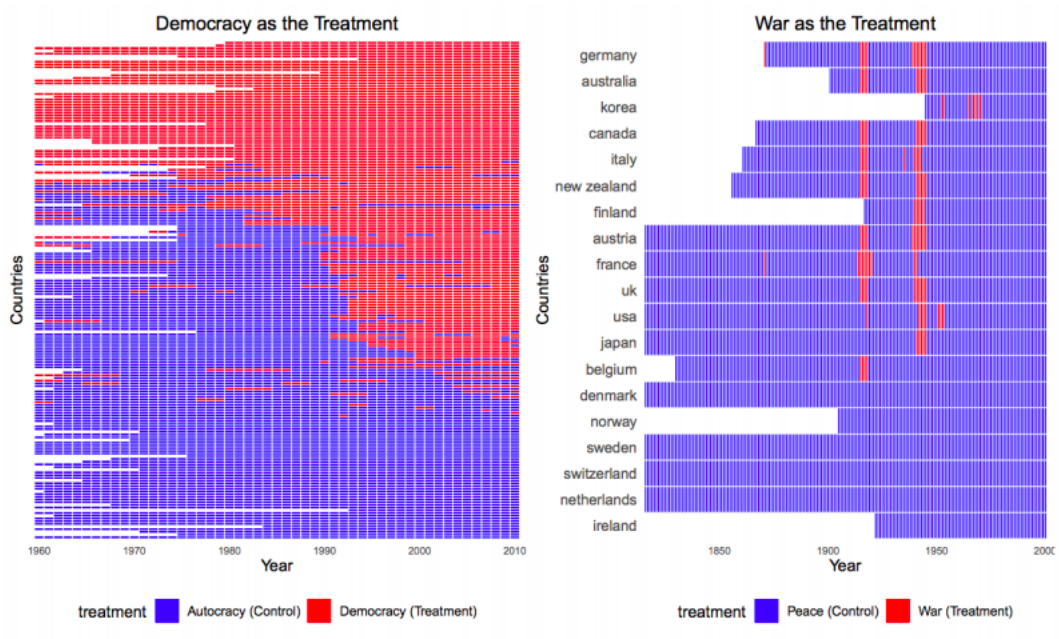


Figure 1: **The Treatment Variation Plots for Visualizing the Distribution of Treatment across Space and Time**. The left panel displays the spatial-temporal distribution of treatment for the study of democracy's effect on economic development (Acemoglu et al., 2019), in which a red (blue) rectangle represents a treatment (control) country-year observation. A white area represents the years when a country did not exist. The right panel displays the treatment variation plot for the study of war's effect on inheritance taxation (Scheve and Stasavage, 2012).

Visualizing censored data is another important step in exploring event history data. However, with a large data set, it is often hard to do so just by looking at the data. A plot like the one below is one way to work around this problem. This figure provided by Yamaguchi (1991) displays six distinct possibilities for event history data. The solid line with an asterisk at the end indicates an event of interest has taken place. The solid line with the open circle at the end indicates that an event other than the event of interest has ended the period of observation (such as an individual dropping out of a study). T_0, or time open, indicates when study began. If an observation experienced the event of interest before T_0, it is left censored. Likewise, T_1, or time closed, represents the end of study. If an observation experienced the event of interest after T_1, it is right censored.

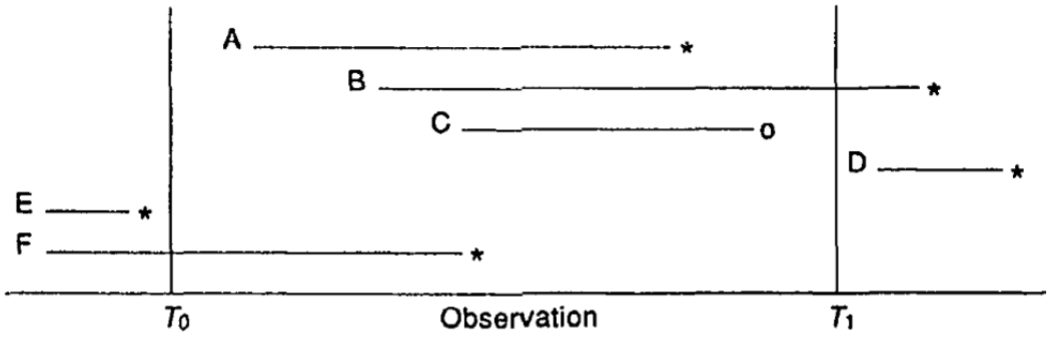


Figure 1.1. Right and Left Censored Observations
NOTE: * = occurrence of event of interest
o = occurrence of event other than event of interest

3 Modeling Event History Processes

To recap, event history analysis typically involves the statistical examination of longitudinal data collected on a set of observations. The dependent variable measures the duration of time that units spend in a state before experiencing some event. In a valuable introduction to this method, especially for social science use, Box-Steffensmeier and Jones (2004) describe how event history data are generated from failure-time processes. A failure-time process

consists of units of analysis observed at some natural starting point, after which they move into some state and are observed over time. A unit, at any given point in the process, is "at risk" of experiencing some event. An event represents a change or transition from one state to another state (for example, losing office in an election). After the event is experienced, the unit is either no longer observed or is at risk of experiencing another kind of event (or returning to the previously occupied state). In some instances, units are not observed experiencing an event, i.e., no transition is made from one state to another. Such cases are treated as censored, because although the event may be experienced, subsequent history after the last observation is unobserved.

The basic logic underlying parametric event history models is to directly model the time dependency exhibited in event history data. This is easily done by specifying a distribution function for the failure times. There are different ways to go about doing this, of which Cox (1972) provides the classic text on event history analysis that models the dependent variable as a hazard ratio.

One application of this modeling approach can be found in Holmaat, Adolph, and Prakash (forthcoming). Here, the hazard of a country joining the Responsible Care program is modeled as a function of t using a Cox proportional hazards model:

$$h_1(t) = h_0(t)exp(x_{i,t-1}\beta + z_{i,t1}\gamma)$$

In this model, $h_i(t)$ is the hazard function for country i and $h_0(t)$ is the baseline hazard function. The covariate of interest $x_{i,t-1}$ measures a specific independent variable tested. $z_{i,t1}$ refers to a vector of time-varying covariates that is controlled for.

In Chapter 9 of *Bankers, Bureaucrats and Central Bank Politics: The Myth of Neutrality* (2013), Adolph uses a Cox proportional hazards model to show the survival of central bankers in office over time. To visualize model output, Adolph presents the following graph of the estimated survival function for mean central bankers, to show the expected median survival in years. This displays how the outcome variable behaves over time.

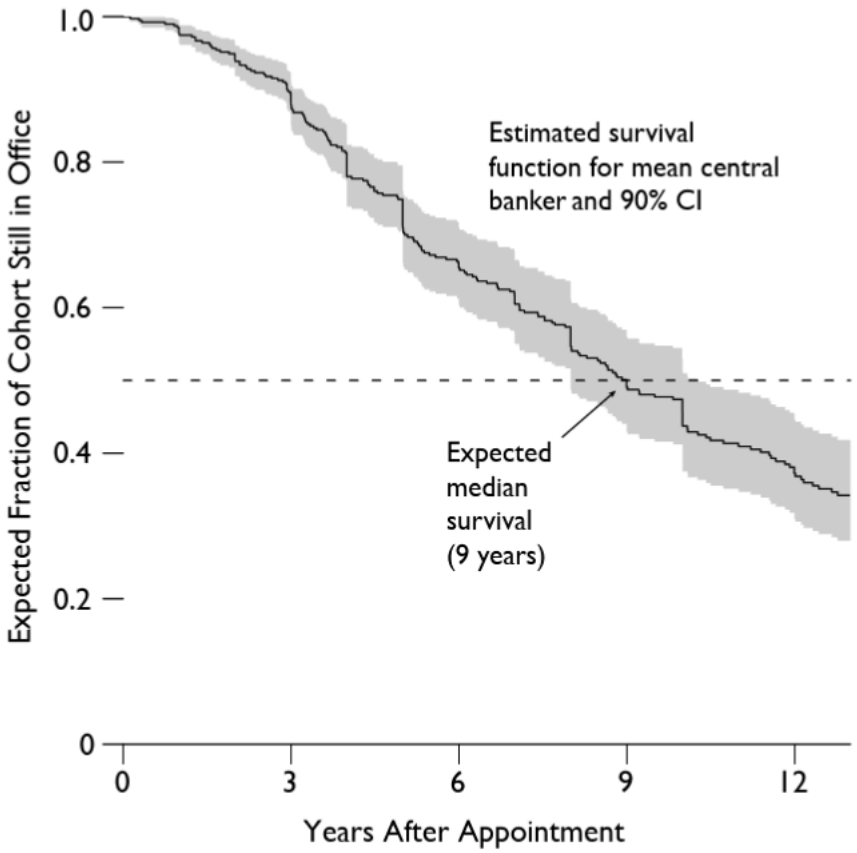
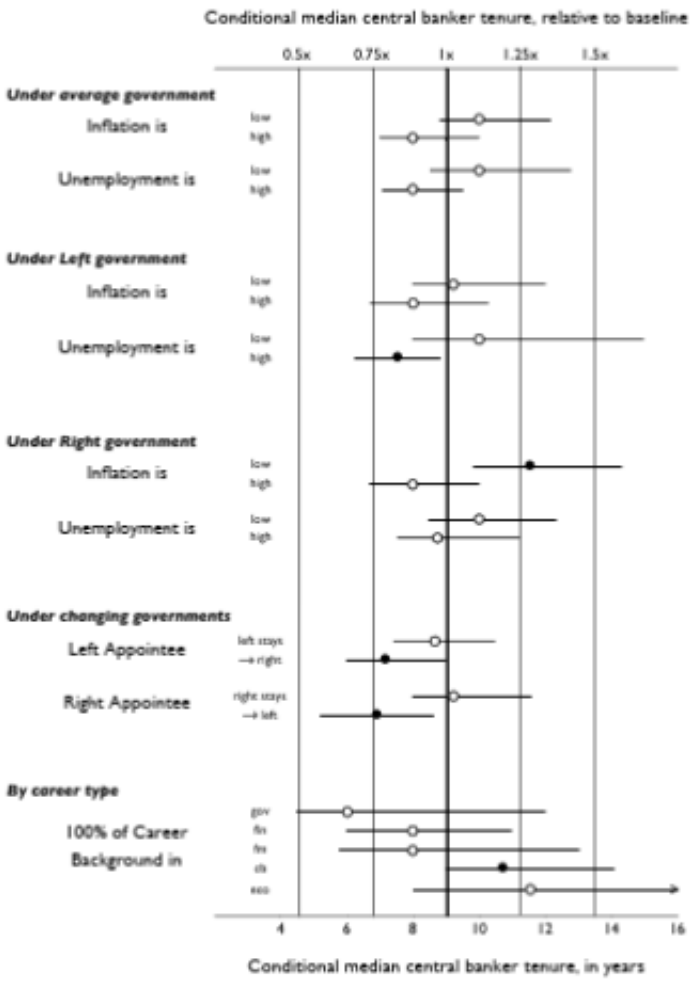


Figure 9.6. *The survival curve for central bankers estimated at the means of the covariates.* The horizontal line marks the point at which half the entering cohort has left the central bank, known as the expected median tenure or halflife. The sharp vertical jumps in the survival curve are clustered at yearly intervals.

Further, Adolph adds to this baseline model output using a ropeladder graph, by varying key explanatory variables and displaying the changes in the dependent variable – here the expected years surviving in office. The plot is displayed below.

These visualizations offer a great example of the effective combination of two different visualization techniques to communicate a model outcome. Adolph first displays how the dependent variable of interest varies over time on average, and then uses a different visualization technique to communicate how this outcome changes when key independent variables are changed (either increasing them by one standard deviation, or going from the 25^{th} percentile to the 75^{th} percentile of the independent variable distribution).



4 Conclusion

Event history data offers a valuable way to describe, explain, and predict the occurrence of events. However, analyses of event history data could often rely on confusing terms and concepts, further exacerbated by nonintuitive visualizations. Here we have outlined strategies for making event history terminology and visualizations more accessible to readers. Next, we discussed the visual means of uncovering patterns in event history analysis when conducting exploratory data analysis. Lastly, we outlined the broad ways to visualize event history models, identifying the primary sources that serve as valuable guides in event history analysis and visualization.

References

Adolph, C. (2013). *Bankers, bureaucrats, and central bank politics: The myth of neutrality.* Cambridge University Press.

Box-Steffensmeier, J., & Jones, Bradford S. (2004). *Event history modeling a guide for social scientists (Analytical methods for social research).* New York: Cambridge University Press.

Cox, D. (1972). Regression Models and Life-Tables. *Journal of the Royal Statistical Society. Series B (Methodological),* 34(2): 187-220.

Holtmaat, E. A., Adolph, C., Prakash, A. (forthcoming). The Global Diffusion of Environmental Clubs: How Pressure from Importing Countries Supports The Chemical Industry's Responsible Care Program. *World Development.*

Imai, K., Kim, I.S., & Wang, E. (2020). *Matching Methods for Causal Inference with Time-Series Cross-Sectional Data.* Working paper presented at the 2018 Midwest Political Science Association Annual Meeting.

Mills, M. (2014). *Introducing survival and event history analysis.* London: Sage Publications.

Syse, A., Dzamarija, M. T., Kumar, B. N., & Diaz, E. (2018). An observational study of immigrant mortality differences in Norway by reason for migration, length of stay and characteristics of sending countries. *BMC Public Health,* 18(1), 1–12.

Yamaguchi, K. (1991). *Event History Analysis.* London: Sage Publications.