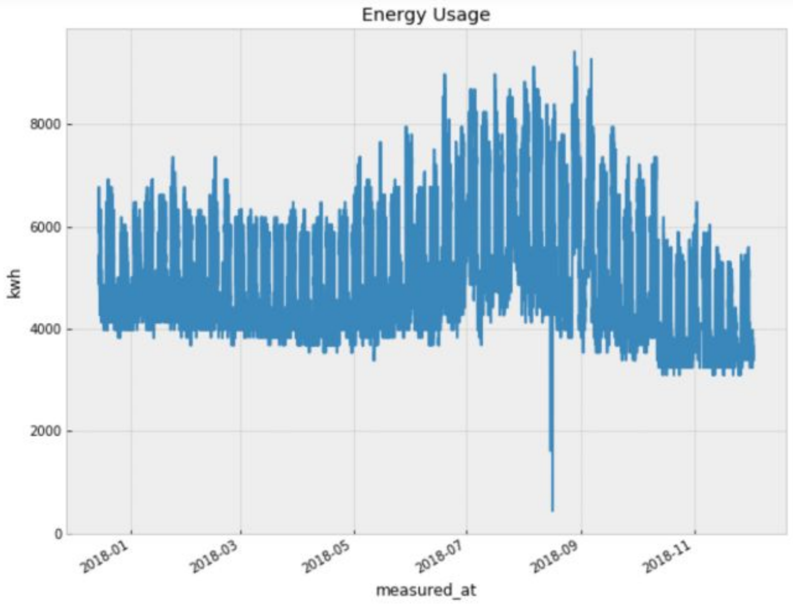
CSSS569 Group project

Visualizing Time Series Data

Breon Haskett, Ivy Ding, Wei Xia, Dorothy Mangale, David Whitlock, Jacqueline Zhou, Sung H. Park, and William Atienza.

Overview:

The utility of time trend data for decision-making is unquestionable, however, displays of these data are often visually unappealing and challenging to decipher beyond very obvious patterns. Seasonality and structural breaks can be easy to pick out, particularly if the data set is simple, and the time scale spans over a short period of time. So how should one approach visualizing time series data? Time series trends of continuous variables are often displayed in monochrome, including applying different line styles to distinguish between groups in the data. However, this should be done with thoughtful consideration of how line color and type might interact to cause cognitive issues. Classic time series trends in scientific publications tend to be boring and are published with minimal labeling and text. This contrasts with the practice or art of presenting figures or tables that speak for themselves. An example of the former is included below.



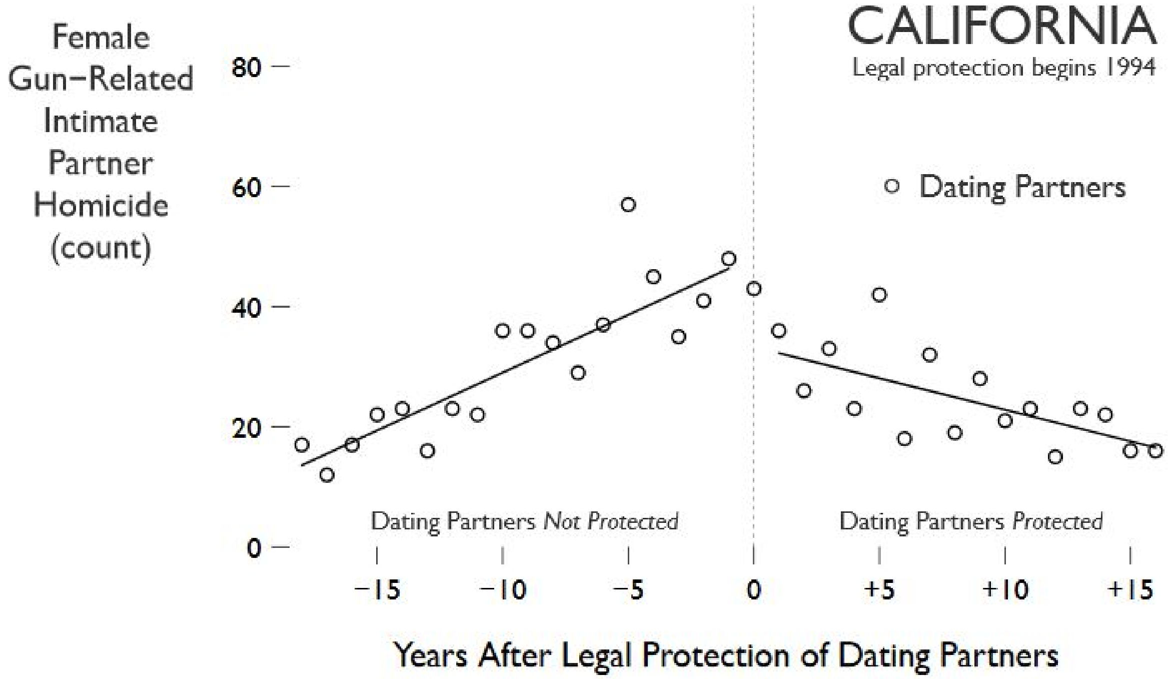
Aesthetics aside, the most important consideration must be the purpose of the graphic. What information is it meant to be sharing? This drives how a central research question is formulated, how it will be answered, and what kinds of data or data transformations will be necessary. Is it of interest to plot mean values or individual level data? Does the time series data need to be a planar figure or would the best use of data be to include a third dimension? How many figures will one need to pass the message across? Is it necessary to plot all data across all observed times or would focussing on areas where the data is unexpected be a better approach? Below we summarize key lessons learned from our group discussion.

Lessons learned

1) Presenting the time variable

Time series data could be seemingly easy to plot. However, deliberate consideration is always necessary before deciding the specific kinds of graphs we use and the way we present each variable. The most prevalent and intuitive way for people to present the time variable is by putting it on the x-axis. While it seemingly easy at first, an important lesson we learned from the class is that there is actually a lot to consider in this process. For instance, it is essential to choose the right scale of time for maximizing readability. This closely relates to the research questions, as researchers need to figure out what they are going to emphasize before visualizing the data. One should note that different time scales might lead to different interpretations of the same piece of data, which researchers need to be cautious of. Moreover, it is wise to present the time series logarithmically when having more events to be visualized closer to one end, which will expand the portion of the x-axis where most relevant and more accurately reflecting the distribution without improperly skewing it to any one extreme. Nonetheless, we need to ensure the labels to be easily interpretable when doing this. Finally, small multiples could be employed to allow distinct contrast among different time periods (i.e. years, months, data collection waves, and so forth).

Legibility and user-friendliness are always important considerations when presenting the time variable, but as researchers we can rely on some notions of the audience understanding of x-axis time variables to our epistemological advantage. For instance, an intervention in your data may occur at a particular time. While it is common to demonstrate this information as a vertical line intersecting your x-axis time variable, perhaps counting down until the intervention, and then up from the intervention could orient audience thinking around the meaning of the intervention. See the following figure as an example from Alvarado et al (2015):



Best practices for time series data

1) Use interactive time series plots with the prospective audience in mind.

The New York Times graphics, as seen in our classes and in books, has some very elaborate time series designs. However, one must be careful about creating such visualization since certain audience may not perceive such design as interesting. For instance, only 10-15% of the online visitors of the New York Times interactive graphics are clicking and utilizing the interaction function embedded in the website (Nediger, 2018). This means that 85% of the readers are not inclined enough to go through the process of clicking and dragging icons or tabs to understand the data. Thus, unless the interactive functions are pivotal in explaining the phenomenon/data of interest, there is no need to go the extra mile just to find out that people are not exploring all the functions and thus aren't getting the entire picture of what is happening.

2) Be aware of the fact that the time interval between each tick and/or how long of a time period is being portrayed can impact the way that the audience interprets the results in time series data graphs.

As seen below in Exhibit 1, time series data interpretation is very sensitive to the time interval between each tick or the period of time depicted in the graph. A steep incline (see picture 1. In Exhibit 1.) can actually turn out to be a small increase when having seen the graph from a more zoomed-out perspective (see picture 4. In Exhibit 1.). Hence, when creating such graph, one needs to be aware of the distortion that one is creating. In this case, either presenting small multiples (as seen in Exhibit 1.) or creating an interactive graph that enables the reader to zoom out or in can help.

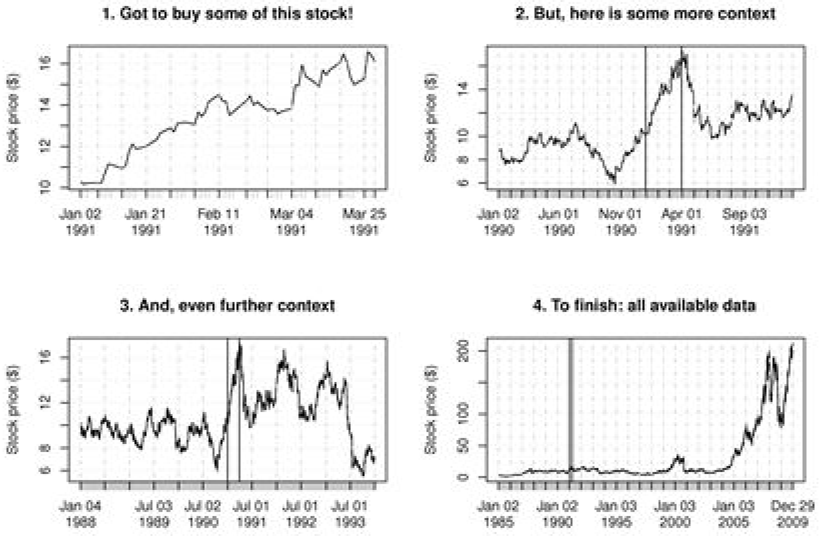


Exhibit 1. How difference in time intervals create different visuals. Source: Dunn (2019)

3) Be aware (and adjust for) changes that take place over time before graphing/visualizing time series data.

Many variables change over time. The change in absolute value is not a problem because this is what we are actually interested in. However, there are things that change not in absolute terms but in relative terms become a problem when not addressed properly in graphs. For instance, many monetary factors are impacted by inflation of the currency or the time of measurement. $1000 dollars in 2002 will not be of the same value in 2020. Thus, when creating visuals that entail such factors, the graphs need to be adjusted for such changes that cannot be detected by readers. Otherwise, the interpretation can be misleading.

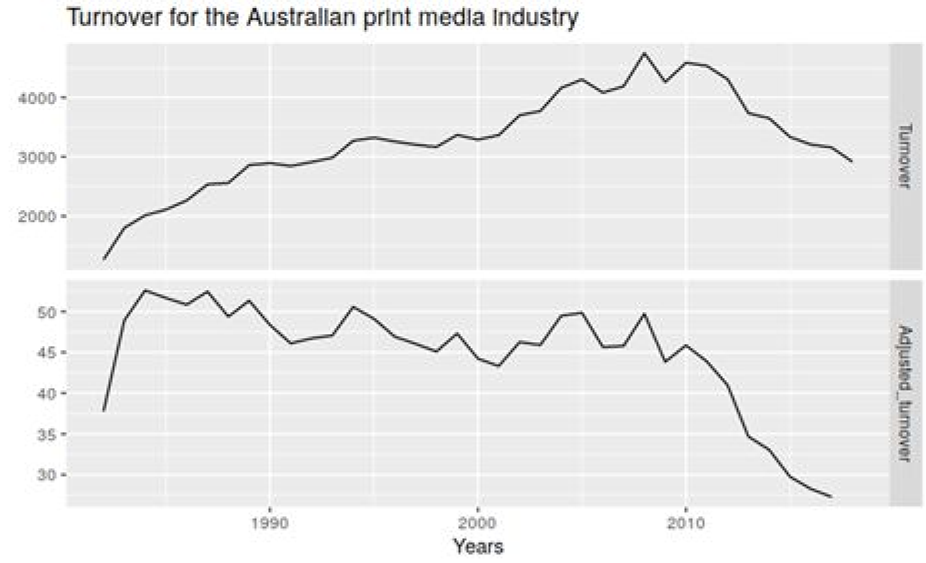


Exhibit 2. Difference between inflation-adjusted and not adjusted graph over time Source: Hyndman and Athanasopoulos (2018)

4) Regarding the use of spaghetti plots

“Spaghetti” plots refer to time series plots with multiple lines that each represent the trend of repeated measures of a variable for each subject across time. These plots can be informative when the number of subjects is low, but as the number of subjects grows, they quickly become difficult to parse. A situation in which spaghetti plots are useful is when we want to highlight certain outliers or exceptions to the general trend. For example, Adolph, Quince, and Prakash (2016) (Exhibit 3, left) highlight certain countries of interest in their use of spaghetti plots by labelling those countries in the plot. When we use spaghetti plots, we are often interested in discussing the information contained in them at the aggregate level as well, so it is good practice to also show the information of interest plotted at the aggregate level as Adolph, Quince, and Prakash (2016) do (Exhibit 3, right).

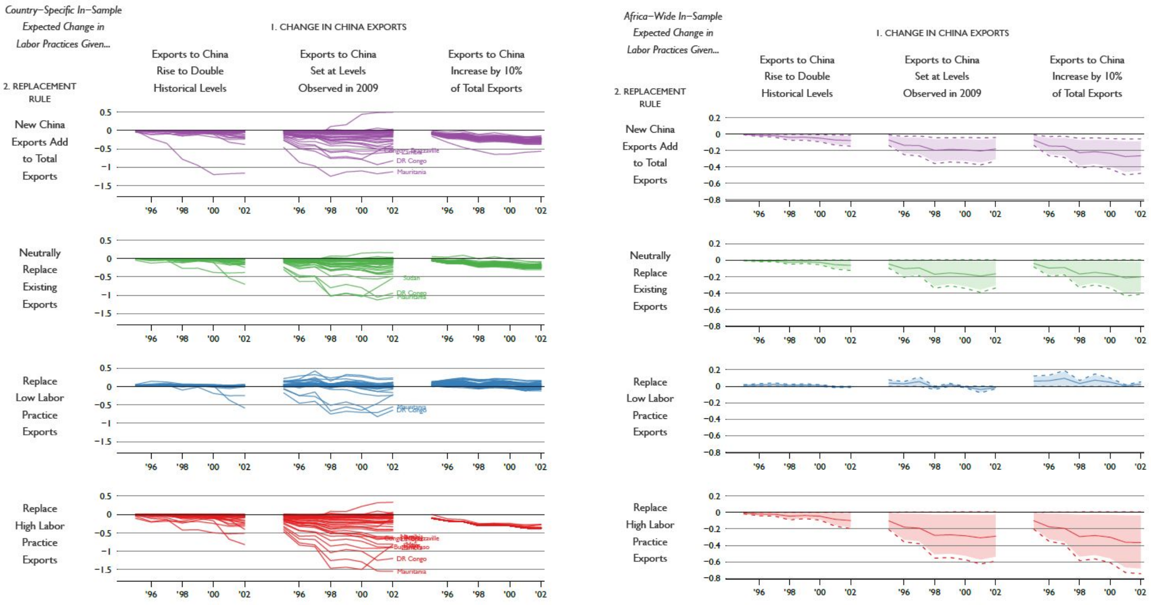
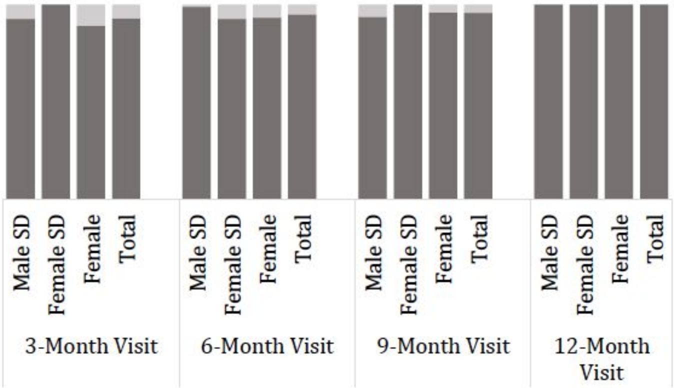


Exhibit 3. In-sample counterfactual change in African labor practice (Mosley-Uno) under higher levels of China exports, by country, year, and replacement rule. (Left) Continent-wide weighted average in-sample counterfactual change in African labor practice (Mosley-Uno) under higher levels of China exports, by year and replacement rule. (Right) Source: Adolph, Quince, and Prakash (2016), Figure 4 and 5

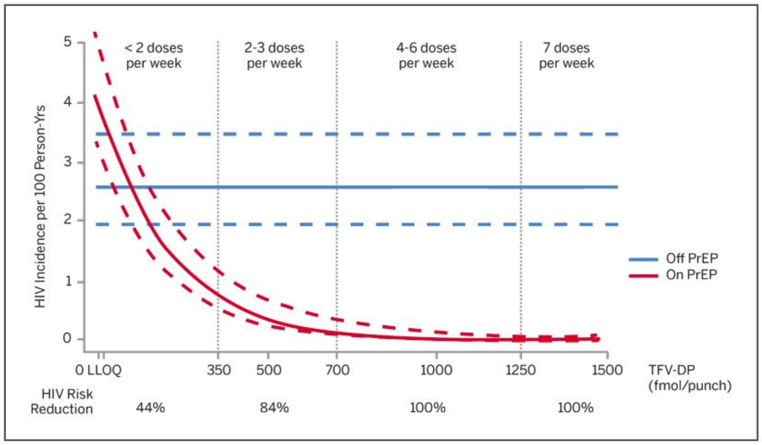
5. When there’s too much data

In the below example from the HIV research world, the bar graph below had been proposed as a visual to explore trends in pre-exposure prophylaxis (PrEP) adherence of individuals in HIV sero-discordant couples receiving one of two types of HIV self-test kits. These data are collected over a period of one year at baseline and four different time groups for one arm of the study, and only two follow up visits post-baseline for the other two arms. Although this figure is easy to create, we are unable to show trends at the individual level, which would have more relevance from a clinical standpoint.

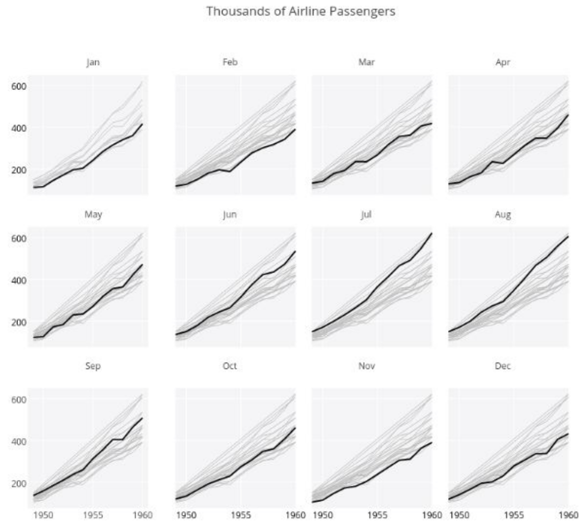


We need to display changes over time, but this visual doesn’t suffice. Additionally, there are questions about whether to display individual data (there are more than 200 individuals in the data set) or if obtaining presenting mean adherence numbers by study arm over time would be better. Policy audiences tend to prefer simple and distilled information, with clear take-home points. Although this might be at the expense of the actual variation in the data. The example from the Lancet shows how most PrEP adherence data from observational studies is typically presented.

FIGURE 1. Adherence Level Needed for Full PrEP Benefit

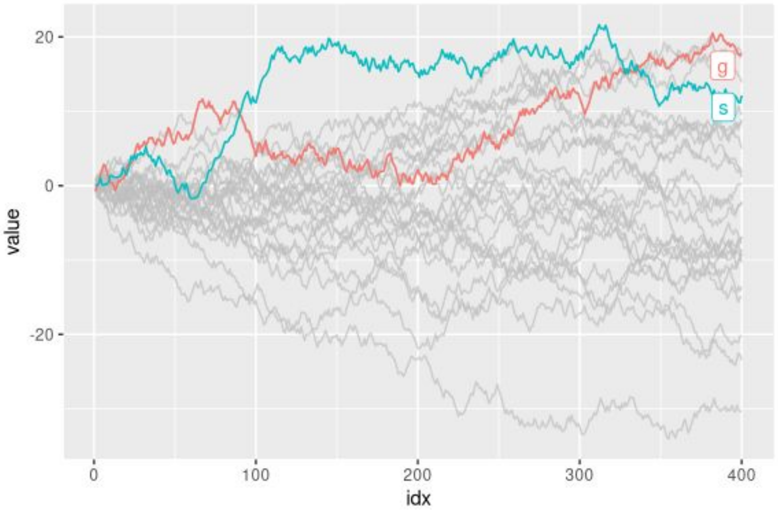


Continuous data are easy to plot for as long as the data is available. Challenges arise when there are multiple groups for comparison and trend lines cross each other. Additionally, if there are multiple subgroups represented in the data, the graph may end up looking cluttered and hard to decipher. Complex analyses in general could make producing simple but information-rich figures challenging. A potential solution would be to 1) incorporate small multiples and create different panels for each group, with mean adherence over time, and include the range of the data. For example:



Another way to deal with large amounts of parallel time series data is to only highlight interesting cases. Problems often arise when labels on line plots are applied too liberally, or when too many color codes exist simultaneously. A useful package when plotting time series data with many factors in R is gghighlight, which can selectively apply an aesthetic to factors which meet a specific condition. A second solution when lookup of individual values need not be as precise as in a line plot is to use a time-scale heatmap. Rows of factors in the time-series heatmap may be sorted after the fact to produce patterns that may be difficult to identify in a series of lines.

Visualizing Time Series Data

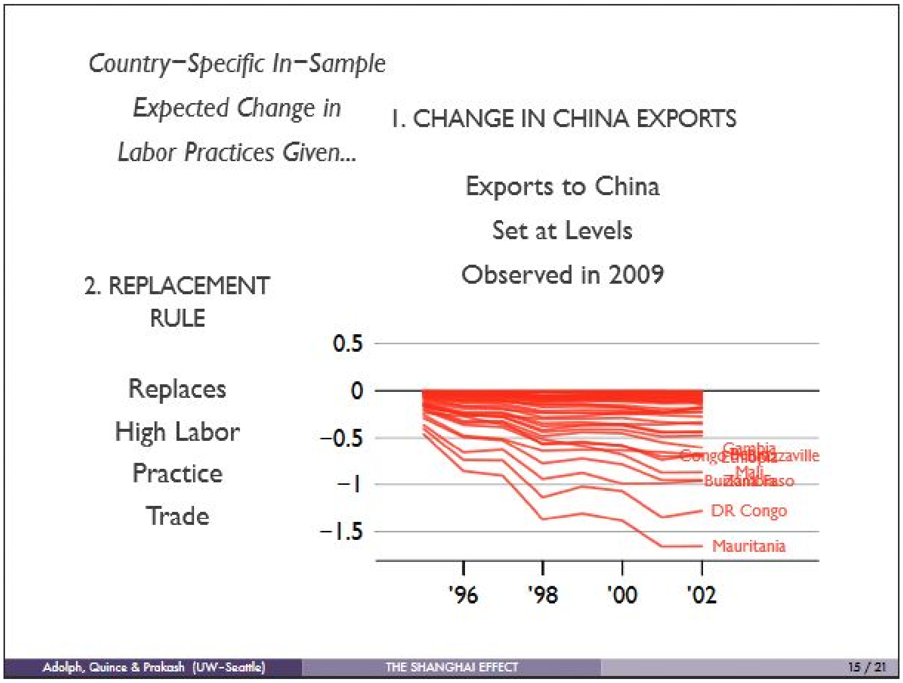


Outstanding Problems

1) Do you actually need several time series plots?

There is the possibility that due to the nature of one's data and the time dimension available that it seems valuable to have many time series plots in a presentation. Large grids of small multiples per many countries, or a list of different outcomes across time are common, but are they substantively valuable to the presentation? We spoke in our group for some time about the time series as an intermediary step. It can be enticing because of the ease of use plotting an x-axis time variable to ask the audience to review many plots, however time series plots can often invite more questions which may be more readily answered by disaggregating some information developed in the time-series plot.

Visualizing Time Series Data



Country–Specific In–Sample Expected Change in Labor Practices Given...

1. CHANGE IN CHINA EXPORTS

Exports to China Set at Levels Observed in 2009

2. REPLACEMENT RULE

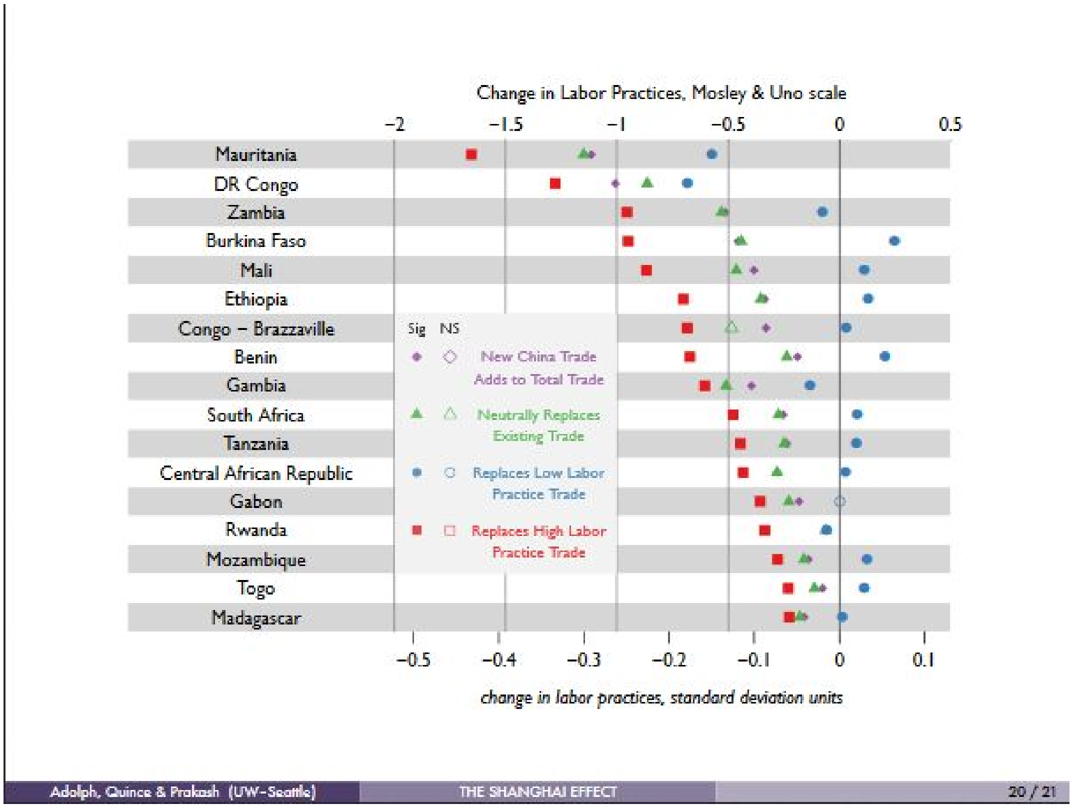
Replaces High Labor Practice Trade

Adolph, Quince & Prakash (UW–Seattle)

THE SHANGHAI EFFECT

15 / 21

From this chart one may begin to ask questions about the unnamed countries here, or the meaning of this time series in a slower-paced way. Great substantive value can be found if one can resist the urge to plot time-series data as an x-axis time series plot. Chris reorients some of this example information in a follow up plot:



Hyndman, R. J., & Athanasopoulos, G. (2018). Forecasting: principles and practice. OTexts.

Nediger, M. (2018, July 17). 10 Data Visualization Best Practices for the Web. Retrieved from https://www.webdesignerdepot.com/2018/07/10-data-visualization-best-practices-for-the-web/

Dunn, K. (2019). Process Improvement Using Data. Retrieved from https://www.citationmachine.net/apa/cite-a-website/search?utf8=✓&q=https://learnche.org/pid/data-visualization/time-series-plots

References

1. Grant RM, Anderson PL, McMahan V, et al. Uptake of pre-exposure prophylaxis, sexual practices, and HIV incidence in men and transgender women who have sex with men: a cohort study. *Lancet Infect Dis*. 2014;14(9):820-829

Adolph, C., Quince, V., and Prakash, A. (2016). The Shanghai Effect: Do Exports to China Affect Labor Practices in Africa? *World Development*, forthcoming.

Hyndman, R. J., & Athanasopoulos, G. (2018). Forecasting: principles and practice. OTexts.

Nediger, M. (2018). 10 Data Visualization Best Practices for the Web. Retrieved from https://www.webdesignerdepot.com/2018/07/10-data-visualization-best-practices-for-the-web/

Dunn, K. (2019). Process Improvement Using Data. Retrieved from https://www.citationmachine.net/apa/cite-a-website/search?utf8=✓&q=https://learnche.org/pid/data-visualization/time-series-plots