Package Demo: emdatr

Gopi Goteti

May 12, 2014

This vignette first provides an overview of the EMDAT database[1] and discusses some of the issues with EMDAT data - particularly, lack of entire data accessibility, static and inconsistent summary reports and the lack of auxiliary financial and demographic data. This is followed by a description of the R package *emdatr* and how it address some of the above issues with EMDAT data. The use of *emdatr* is demonstrated, followed by the duplication of summary graphics presented in the one of EMDAT's recent publications.

1 Overview of EMDAT Database

The International Disaster Database, EMDAT from the Center for Research on the Epidemiology of Disasters (CRED, Belgium) is often used as a reference for losses on human life and property resulting from natural and man-made disasters. This database has over 20,000 country-level records from the early 1900s to the present. Data is available for free from EMDAT.

Some issues with EMDAT data are as follows.

- **Data Inaccessibility**
  - EMDAT provides only partial information on the geographical extent of the disaster. Country information is always provided, but the specific provinces and sub-provinces within a country are not provided through the the website. The region field displayed on the website typically includes a couple of provinces followed by "...".
  - It appears that access to the entire database is restricted and it is unclear why EMDAT does not release their entire database.
- **Static and Inconsistent Summary Reports**
  - Annual reports published by EMDAT[2] are inconsistent with one another in terms of number of disasters per year or the total number of people affected or killed. For instance, number of disasters in 2002

---

[1] http://www.emdat.be/database

[2] Annual Disaster Statistical Review (ADSR) Reports for 2008 through 2012 were obtained from http://www.emdat.be/publications

    were reported to be 428 in the Annual Disaster Statistical Review (ADSR) report for 2012. But the same number in the 2011, 2010, 2009 and 2008 reports is 421, 421, 422 and 421, respectively!
  - The above issue could partly be due to the static nature of these reports. Whereas data gets updated in the database, the reports generated in the past are not. In the generation of "Web 2.0", a dynamic summary reporting site is reasonable to expect.

- **Data Conventions**
  - Country names used by EMDAT are not always the same as those used by ISO 3166 convention[3]. This issue is relevant when making spatial maps using R.

- **Lack of Auxiliary Information**
  - Financial losses reported by EMDAT are from the year of occurrence of the disaster and are not adjusted for inflation.
  - Annual GDP and population data are often used to project (or "normalize)" historical monetary losses to the present[4]. EMDAT does not provide such information.

2 R Package *emdatr*

The R package *emdatr* addresses some of the above-mentioned issues with the EMDAT data. The goal of the package is to promote the use of EMDAT data, bring transparency to the data, shed light on the limitations of the data, and make the analysis of the data easier through the R language.

2.1 Cleaned and Enhanced EMDAT Data

Raw data was obtained from the EMDAT website and was cleaned, formatted and enhanced. Following is an overview of this procedure.

- **Typographical errors** in country names and disaster types were corrected.
- **ISO 3166 convention** - Country names from EMDAT were mapped to the ISO names by visually comparing the names. The mismatch is names was either due to abbreviations used by EMDAT, for instance - Is for Islands, or anglicized spelling used by ISO. Some countries could not be assigned an ISO name due to geographical splits. Hence, the former countries of Czechoslovakia, Yugoslavia, Serbia Montenegro and Soviet Union have been assigned an ISO name of X_X.

---

[3] http://en.wikipedia.org/wiki/ISO_3166

[4] For instance,

- **GDP and population** data from the World Bank[5] was added, when available, to each of the EMDAT events. Some country codes in the World Bank data have also been found to be inconsistent with ISO 3166 convention. Hence, ROM, PSE, TMP, ZAR were assigned the codes of ROU, WBG, TLS, COD, respectively.
- EMDAT's financial losses are always reported in USA Dollars from the year of occurrence of the disaster. Adjustment of historical losses for inflation requires Consumer Price Index (CPI). The USA **CPI from the Bureau of Labor Statistics**[6] is used in the package to adjust for inflation.

2.2 Getting the Data

After installing the package, load the package along with RCurl (for data extraction from bitbucket.org), ggplot (for graphics) and plyr (for data manipulation).

```
> require(emdatr)
> require(RCurl)
> require(ggplot2)
> require(plyr)
```

The single main function provided by *emdatr* is *extract_emdat*. This could be used to extract a sample of the EMDAT data (which comes with this package) or the entire data. First, load the sample data that comes with the package.

```
> losses_2013 <- extract_emdat()
> dim(losses_2013)

[1] 545  18

> head(losses_2013)

         Start        End     Country                      Location
200 24/04/2013 24/04/2013 Afghanistan Kameh, Dehbala, Lalpur, S ...
201  10/8/2013 14/08/2013 Afghanistan Chakardar, Chak, Jaghatu, ...
202   1/8/2013   7/8/2013 Afghanistan Kabul, Khost, Kunar, Pakt ...
203 25/04/2013 29/04/2013 Afghanistan  Baghlan, Ghor, Balkh pro ...
204   4/2/2013  10/2/2013 Afghanistan  Hirat, Parwan, Kandahar, ...
205 15/09/2013 15/09/2013 Afghanistan Ruyi Du Ab district (Sama ...
                            Type                     SubType      Name Killed
200 earthquake (seismic activity) earthquake (ground shaking)            18
201                         flood               general flood            31
202                         flood               general flood            52
203                         flood               general flood            20
```

---

[5] http://databank.worldbank.org/data/home.aspx

[6] http://www.bls.gov/cpi/tables.htm

```
204                          flood              general flood             10
205            industrial accident                      collapse Coal mine     28
    TotAffected EstDamage     DisNo         Group Year ISO_alpha3   ISO_cntry
200        3531        NA 2013-0151   geophysical 2013        AFG Afghanistan
201          NA        NA 2013-0343  hydrological 2013        AFG Afghanistan
202        2597        NA 2013-0279  hydrological 2013        AFG Afghanistan
203        9500        NA 2013-0178  hydrological 2013        AFG Afghanistan
204        5000        NA 2013-0148  hydrological 2013        AFG Afghanistan
205          17        NA 2013-0359 technological 2013        AFG Afghanistan
    region Pop GDP
200   Asia  NA  NA
201   Asia  NA  NA
202   Asia  NA  NA
203   Asia  NA  NA
204   Asia  NA  NA
205   Asia  NA  NA
```

The default options of *extract_emdat* do not make any adjustments for inflation. Next, obtain the entire dataset with the *inflation* option enabled. This might take a few seconds. The result is that all historical financial losses are adjusted for inflation resulting in equivalent dollar amounts in 2013. If a different year of adjustment is desired, change the *base_year* accordingly.

```
> losses_all <- extract_emdat(sample_only = FALSE, inflation = TRUE)
> dim(losses_all)
```

```
[1] 20854    19
```

```
> head(losses_all)
```

```
       Start        End     Country        Location
1  10/6/1954  10/6/1954 Afghanistan    North Region
2  10/6/1956  10/6/1956 Afghanistan           Kabul
3 00/07/1956 00/07/1956 Afghanistan
4 00/04/1963 00/04/1963 Afghanistan
5  12/6/1964  12/6/1964 Afghanistan          Karkar
6 00/01/1969 00/00/1969 Afghanistan Paktia province
                          Type                        SubType Name Killed
1 earthquake (seismic activity) earthquake (ground shaking)       2000
2 earthquake (seismic activity) earthquake (ground shaking)        100
3                         flood                                     51
4                         flood                                    107
5           industrial accident                  explosion Mine     74
6                       drought                         drought      NA
  TotAffected EstDamage     DisNo         Group Year ISO_alpha3   ISO_cntry
1          NA        NA 1954-0009    geophysical 1954        AFG Afghanistan
2        2000      25.0 1956-0008    geophysical 1956        AFG Afghanistan
```

```
3          NA        NA 1956-0039   hydrological 1956        AFG Afghanistan
4          NA        NA 1963-0065   hydrological 1963        AFG Afghanistan
5         400        NA 1964-0033  technological 1964        AFG Afghanistan
6       48000       0.2 1969-9007 climatological 1969        AFG Afghanistan
  region Pop       GDP Damage_Adjusted_2013
1   Asia  NA        NA                   NA
2   Asia  NA        NA            214.11489
3   Asia  NA        NA                   NA
4   Asia  NA  751.1112                   NA
5   Asia  NA  800.0000                   NA
6   Asia  NA 1408.8889              1.26952
```

All financial losses from EMDAT are reported in Millions of US Dollars. Adjustment for inflation is currently based on the relative ratio of the Consumer Price Index (CPI) of the United States - i.e., the adjustment factor is the ratio of CPI in the *base_year* and the CPI in the year of the disaster. However, such adjustment may be inappropriate since it does not account for any direct economic changes in the country of occurrence. Future updates to the package could incorporate such economic effects.

3 Duplicating Select Graphics from ADSR 2012 Report

Example graphics shown in this section are intended to duplicate those shown in EMDAT's ADSR report from 2012[7]. Graphics shown in this section represent the unique set of charts and graphs shown in the ADSR 2012 report and not the entire set of graphics.

From the entire dataset, identify natural disasters only.

```
> nat_data <- losses_all[losses_all$Group %in% c("climatological", "geophysical",
+     "hydrological", "meteorological"), ]
> nat_data <- droplevels(nat_data)
> # assign missing value to 0s before using cbind in aggregate
> nat_data$Killed[is.na(nat_data$Killed)] <- 0
> nat_data$TotAffected[is.na(nat_data$TotAffected)] <- 0
> nat_data$Year <- as.factor(nat_data$Year)
```

3.1 Figure 1, ADSR Report 2012

Identify number killed and affected per year from 1990 through 2012.

```
> gfx_deaths <- aggregate(cbind(Killed, TotAffected) ~ Year, data = nat_data,
+     FUN = sum)
```

---

[7] Guha-Sapir D, Hoyois Ph., Below. R. Annual Disaster Statistical Review 2012: The Numbers and Trends. Brussels: CRED; 2013., http://www.emdat.be/publications

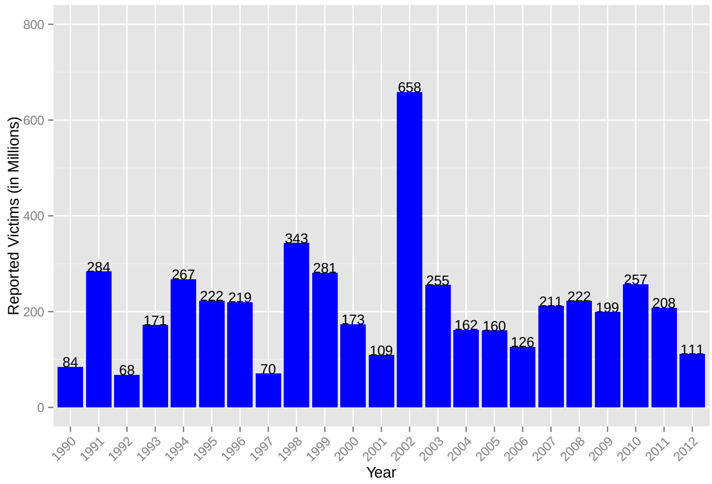

Figure 1: Trends in Victims, Millions, Sum of Killed and Total Affected. Compare with Figure 1, pg. 3 of the ADSR report from 2012.

```
> # total in millions
> gfx_deaths$Total <- (gfx_deaths$Killed + gfx_deaths$TotAffected)/10^6
> gfx_deaths <- gfx_deaths[, c("Year", "Total")]
> gfx_deaths <- gfx_deaths[gfx_deaths$Year %in% seq(1990, 2012), ]
> gfx_deaths <- droplevels(gfx_deaths)
```

Plot number killed or affected by year, similar to the barplot in EMDAT's ADSR report from 2012 (Figure 1, pg. 3 of the ADSR report). See Figure 1.

```
> gfx_bar <- ggplot(gfx_deaths, aes(x = Year, y = Total))
> gfx_bar <- gfx_bar + geom_bar(position = "dodge", stat = "identity", fill = "blue")
> gfx_bar <- gfx_bar + ylab("Reported Victims (in Millions)")
> gfx_bar <- gfx_bar + ylim(0, 800)
> gfx_bar <- gfx_bar + theme(axis.text.x = element_text(angle = 45, hjust = 1))
> gfx_bar <- gfx_bar + geom_text(aes(label = round(Total), hjust = 0.5, vjust = 0),
+     size = 4)
```

Number of events per year from 1990 through 2012.

```
> gfx_events <- as.data.frame(table(nat_data$Year), stringsAsFactors = FALSE)
> colnames(gfx_events) <- c("Year", "Total_Events")
> gfx_events <- gfx_events[gfx_events$Year >= 1990 & gfx_events$Year <= 2012, ]
> gfx_events[gfx_events$Year == 2002, ]
```

```
    Year Total_Events
103 2002          422
```

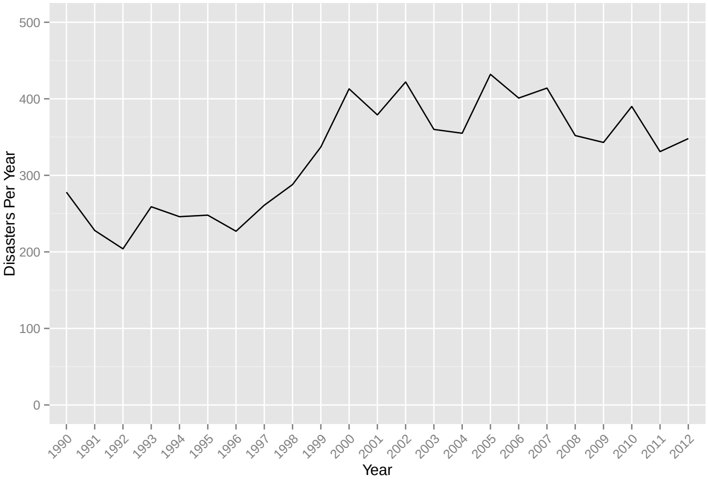

Figure 2: Trends in Disaster Occurrence, EMDAT Reported Disasters Per Year. Compare with Figure 1, pg. 3 of the ADSR report from 2012. Note that the number of events in 2002 were reported to be 428 in the ADSR 2012 report. But the same number in the 2011, 2010, 2009 and 2008 reports is 421, 421, 422 and 421, respectively!

Plot number of events by year, similar to the lineplot in EMDAT's ADSR report 2012 (Figure 1, pg. 3 of the ADSR report). See Figure 2.

```
> gfx_line <- ggplot(gfx_events, aes(x = Year, y = Total_Events, group = 1))
> gfx_line <- gfx_line + geom_line()
> gfx_line <- gfx_line + ylab("Disasters Per Year")
> gfx_line <- gfx_line + ylim(0, 500)
> gfx_line <- gfx_line + theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

3.2 Figure 3 and 6, ADSR Report 2012

In order to replicate the graphic on top 10 countries by loss ((Figure 3 and 6, pg. 15-16 of the ADSR report), a generic function is developed below which could not only be used with loss but also other variables.

```
> Fn_Get_Top_Countries <- function(input_df, var_name, plot_title) {
+     var_vec <- c("Events", "EstDamage", "TotAffected", "Killed")
+     stopifnot(colnames(input_df) == colnames(nat_data))
+     stopifnot(var_name %in% var_vec)
+
+     fun_name <- "sum"
+     if (var_name == "Events") {
+         fun_name <- "length"
+         var_name <- "Year"
+     }
+
+     # summary by country per natural disaster group
+     data_by_group <- aggregate(as.formula(paste(var_name, " ~ ISO_cntry + Group")),
+         data = input_df, FUN = fun_name)
+     colnames(data_by_group) <- c("Country", "Group", var_name)
+
+     # totals by country
+     data_agg <- aggregate(as.formula(paste(var_name, " ~ ISO_cntry")), data = input_df,
+         FUN = fun_name)
+     colnames(data_agg) <- c("Country", "Totals")
+     data_agg <- data_agg[order(data_agg$Totals, decreasing = TRUE), ]
+     cntrys_10 <- data_agg$Country[1:10]
+
+     # merge above two data frames
+     out_df <- merge(data_by_group, data_agg, by = "Country")
+     out_df <- out_df[order(out_df$Totals, decreasing = TRUE), ]
+
+     out_df <- out_df[out_df$Country %in% cntrys_10, ]
+     out_df <- droplevels(out_df)
+
+     out_df$Country <- factor(out_df$Country, levels = rev(cntrys_10))
+     # percentage share
+     out_df$Pers <- out_df[, var_name] * 100/out_df$Totals
+
+     return(out_df)
+ }
```

Use the above function to get natural disaster counts by disaster Group for 2012 for the top 10 countries.

```
> nat_2012 <- nat_data[nat_data$Year == 2012, ]
> nat_2012 <- droplevels(nat_2012)
> gfx_2012_counts <- Fn_Get_Top_Countries(nat_2012, "Events")
> head(gfx_2012_counts, 10)
```

```
          Country          Group Year Totals      Pers
38          China climatological    1     28  3.571429
39          China   hydrological   13     28 46.428571
40          China    geophysical    6     28 21.428571
41          China meteorological    8     28 28.571429
175 United States   hydrological    1     25  4.000000
176 United States climatological    5     25 20.000000
177 United States meteorological   19     25 76.000000
```

```
134   Philippines   hydrological    9     21 42.857143
135   Philippines meteorological    9     21 42.857143
136   Philippines    geophysical    3     21 14.285714
```

Barplot of top 10 countries by number of natural disasters in 2012. See Figure 3.

```
> gfx_bar <- ggplot(gfx_2012_counts, aes(x = Country, y = Year, group = Group))
> gfx_bar <- gfx_bar + geom_bar(aes(fill = Group), position = "stack", stat = "identity")
> gfx_bar <- gfx_bar + ylab("Number of Events") + xlab(NULL)
> gfx_bar <- gfx_bar + coord_flip()
```

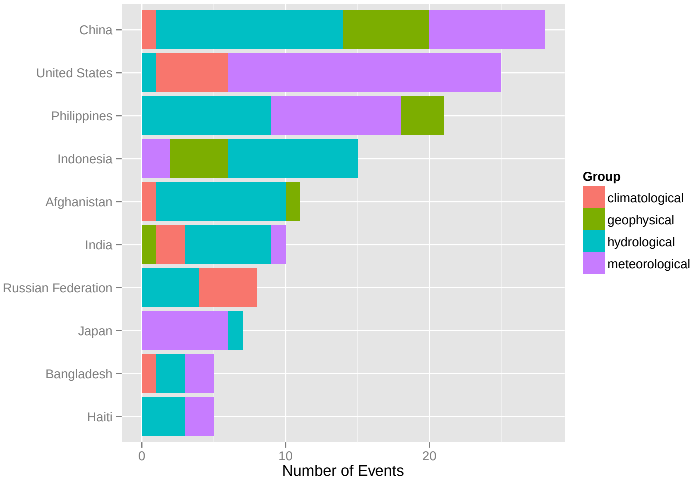

Figure 3: Top 10 countries by number of events in 2012. Compare with Figure 3, pg. 15 of the ADSR report from 2012.

Use the above function to get natural disaster losses by disaster Group for 2012 for the top 10 countries.

```
> gfx_2012_losses <- Fn_Get_Top_Countries(nat_2012, "EstDamage")
> head(gfx_2012_losses, 10)

         Country          Group EstDamage   Totals        Pers
67 United States climatological 20800.000 98469.00 21.12339924
68 United States meteorological 77495.000 98469.00 78.69989540
69 United States   hydrological   174.000 98469.00  0.17670536
12         China climatological    20.200 19754.53  0.10225501
13         China   hydrological 14970.333 19754.53 75.78176108
14         China meteorological  3216.000 19754.53 16.27980778
```

```
15         China    geophysical  1548.000 19754.53  7.83617613
28         Italy climatological  1322.601 17137.60  7.71753876
29         Italy   hydrological    15.000 17137.60  0.08752684
30         Italy    geophysical 15800.000 17137.60 92.19493440
```

Pieplot of these top 10 countries. See Figure 4.

```
> gfx_pie <- ggplot(gfx_2012_losses, aes(x = "", y = Pers, fill = Group))
> gfx_pie <- gfx_pie + facet_wrap(~Country)
> gfx_pie <- gfx_pie + geom_bar(width = 1, stat = "identity")
> gfx_pie <- gfx_pie + coord_polar(theta = "y")
> gfx_pie <- gfx_pie + theme(axis.ticks = element_blank(), axis.text.y = element_blank(),
+     axis.text.x = element_blank())
> gfx_pie <- gfx_pie + xlab("") + ylab("")
```

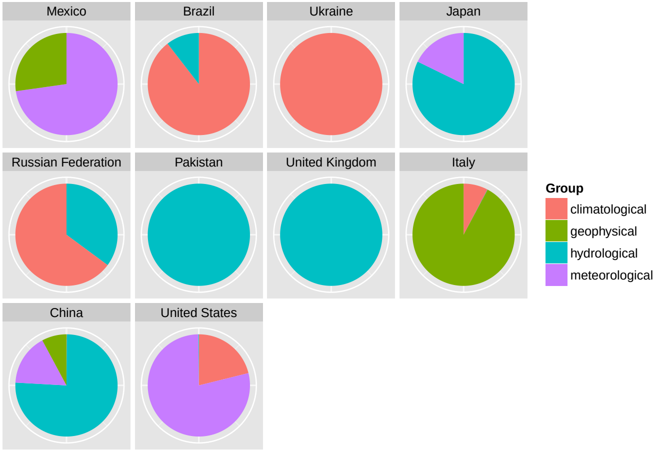

Figure 4: Top 10 countries by losses in 2012. Compare with Figure 6, pg. 16 of the ADSR report from 2012. Note the discrepancies in this plot and the one from ADSR. For instance, Mexico is in the above graphic and not in the ADSR graphic, whereas Philippines is present in the ADSR graphic and not in the above graphic. Also the percentage share of the Group is not always the same between these two graphics.

3.3 Map 3, ADSR Report 2012

In Map 3 of the ADSR Report (see pg. 33) the color scheme of the barplots and the color scheme of the continental regions in the map overlap resulting in a misrepresentation of the summary statistics. Below code reproduces the statistics presented in Map 3.

First, compute the regional disaster losses and the percent share of each region within each Group.

```
> gfx_reg1 <- ddply(nat_2012[, c("EstDamage", "Group", "region")],
+                  .(region, Group),
+                  summarize,
+                  tot_by_group = sum(EstDamage, na.rm = TRUE))
> gfx_reg2 <- ddply(nat_2012[, c("EstDamage", "Group", "region")],
+                  .(Group),
+                  summarize,
+                  tot_by_reg = sum(EstDamage, na.rm = TRUE))
> gfx_reg <- merge(gfx_reg1, gfx_reg2, by = "Group", all.x = TRUE)
> gfx_reg$share <- gfx_reg$tot_by_group * 100 / gfx_reg$tot_by_reg
> head(gfx_reg)
```

```
          Group   region tot_by_group tot_by_reg       share
1 climatological   Africa        0.000   26632.80  0.00000000
2 climatological Americas    22460.000   26632.80 84.33209860
3 climatological     Asia       20.200   26632.80  0.07584632
4 climatological   Europe     4152.601   26632.80 15.59205508
5    geophysical Americas      675.000   18536.31  3.64150068
6    geophysical     Asia     2061.314   18536.31 11.12040938
```

Plot percent share of each region within each Group. See Figure 5

```
> gfx_bar <- ggplot(gfx_reg, aes(x = Group, y = share, group = region))
> gfx_bar <- gfx_bar + geom_bar(aes(fill = Group), position = "dodge", stat = "identity")
> gfx_bar <- gfx_bar + facet_wrap(~region, scales = "free_y")
> gfx_bar <- gfx_bar + ylab("Percent Share") + xlab(NULL)
> gfx_bar <- gfx_bar + theme(axis.text.x = element_blank(), axis.ticks.x = element_blank())
```

4 Maps using *rworldmap*

Make a map of global financial losses from all disasters for 2013.

During the vignette creation process, the following code on making a map resulted in an error, possibly due to formatting errors in the TeX script. Some expertise in TeX is required to resolve this error, but the author does not have it. The below code works on its own but not within the vignette creation process. Hence, the below three chunks of code are not evaluated and are only shown for reference. Future updates to the package would try to fix this error.

First, get the total loss by country using the ISO3 country names.

```
> losses_cntry <- ddply(losses_2013,
+                        .(ISO_alpha3),
```

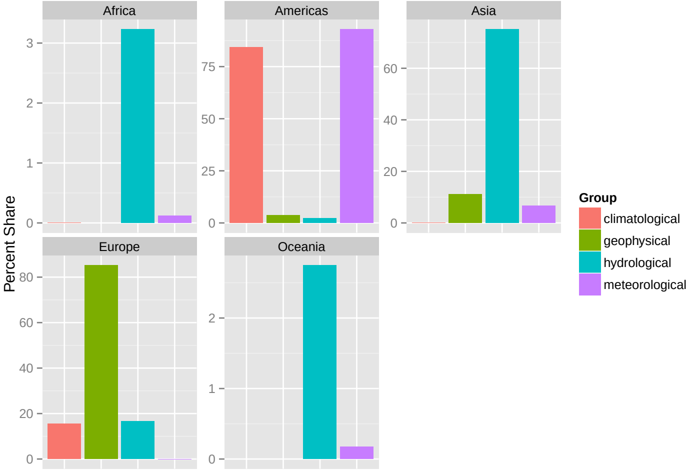

Figure 5: Percent share of disaster losses by disaster group. Compare with Map 3, pg. 33 of the ADSR report from 2012.

```
+                       summarize,
+                       total = sum(EstDamage, na.rm = TRUE))
> # remove "X__X" introduced during the cleaning process
> losses_cntry <- losses_cntry[losses_cntry$ISO_alpha3 != "X__X", ]
> # convert to billions; exclude 0s and NAs
> losses_cntry$total <- losses_cntry$total / 10^3
> losses_cntry <- losses_cntry[!is.na(losses_cntry$total) & losses_cntry$total > 0, ]
> head(losses_cntry)
> summary(losses_cntry)
```

Using the rworldmap package, create a data frame compatible with rworldmap plotting functions.

```
> require(rworldmap)
> losses_cntry <- joinCountryData2Map(losses_cntry,
+                                     joinCode = "ISO3",
+                                     nameJoinColumn = "ISO_alpha3")
> class(losses_cntry)
```

Print the map of losses by country for 2013.

```
> gfx_map <- mapCountryData(losses_cntry,
+                           nameColumnToPlot = "total",
+                           mapTitle = "",
+                           colourPalette = "terrain",
+                              addLegend = FALSE)
> gfx_map <- do.call(addMapLegend,
+                    c(gfx_map,
+                      legendLabels = "all",
+                      legendWidth = 0.3,
+                      sigFigs = 1))
```

5 Summary

The EMDAT database provides valuable information on human and financial losses from natural disasters around the world. Some of the issues with the EMDAT data are lack of entire data accessibility, static and inconsistent summary reports, and the lack of auxiliary financial and demographic data. The *emdatr* package addresses some of these issues. The examples provided in this vignette demonstrate the functionality provided by the *emdatr* package. The goal of the *emdatr* package is to promote the use of EMDAT data, bring transparency to the data, shed light on the limitations of the data, and make the analysis of the data easier through the R language and the plethora of open source packages built around it.