Cartel detection

Margherita Atzei, Sebastian Kimm Friedenberg, Oscar Krumlinde, Filip Mellgren

2020-03-11

Contents

Set hyper-parameters

```
dc <- 0.7 # Decision criteria cut off
```

Data cleaning

Create the necessary aggregate statistics we will be analysing throughout

For convenience, we drop observations where CV or altrd does not exist or are infinite. These two variables provide the bulk of the analysis and dropping them early therefore has no effect on most graphs as they would have been dropped anyway. For the results that do not depend on the two variables, the difference is small and doesn't change the conclusion.

```
# NOTE that some observations disappear here
# Lost observations due to missings: 4434 - 3811 = 623
df_agg_filtered <- df_agg %>% select(project, CV, altrd, contract_type) %>% filter_all(any_vars(!is.inf

df_agg <- left_join(df_agg_filtered, df_agg)
```

Structural screen, descriptive statistics

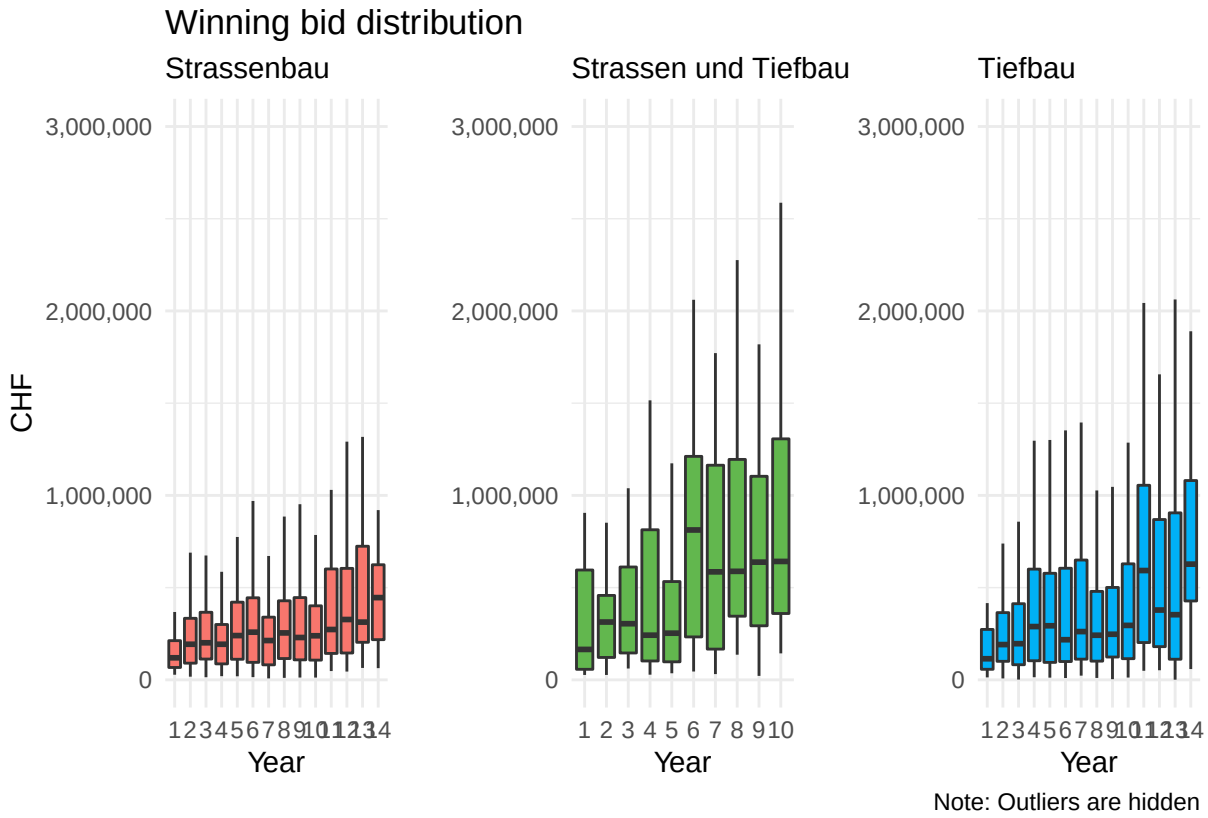
We will be looking at boxplots and line graphs, therefore we create convenience functions for future use:

```
box_plot <- function(df, variable, type, color, type_name, y_min, y_max, note){
  plot <- melt(df, id.var = c("contract_type", "year"), measure.vars = c(variable)) %>%
    filter(contract_type == type) %>%
```

```
    ggplot(aes(x = as.factor(year), y = value)) +
    geom_boxplot(outlier.shape = NA, fill = color) + theme_minimal() +
    labs(y = "", x = "Year", subtitle = type_name, caption = note) +
    coord_cartesian(ylim = c(y_min, y_max))
  return(plot)
}

avg_var_year <- function(df, var){
  p <- df %>% select(project, year, contract_type, var) %>%
    rename(var_ = var) %>%
  group_by(year, contract_type) %>%
  summarise(avg_var = mean(var_)) %>%
  ggplot(aes(x = year, y = avg_var)) +
  geom_line(size = 0.75, aes(fill = as.factor(contract_type))) +
  geom_point(size = 2.5, aes(color=factor(contract_type))) +
  theme_minimal() +
  scale_color_manual(name = "Contract type", values=c("#f8766d", "#62B74E", "#00b0f6"),
                     labels = c("Strassenbau", "Strassen und Tiefbau", "Tiefbau"))
  return(p)
}
```

What is the distribution of the winning bids by category type and year?

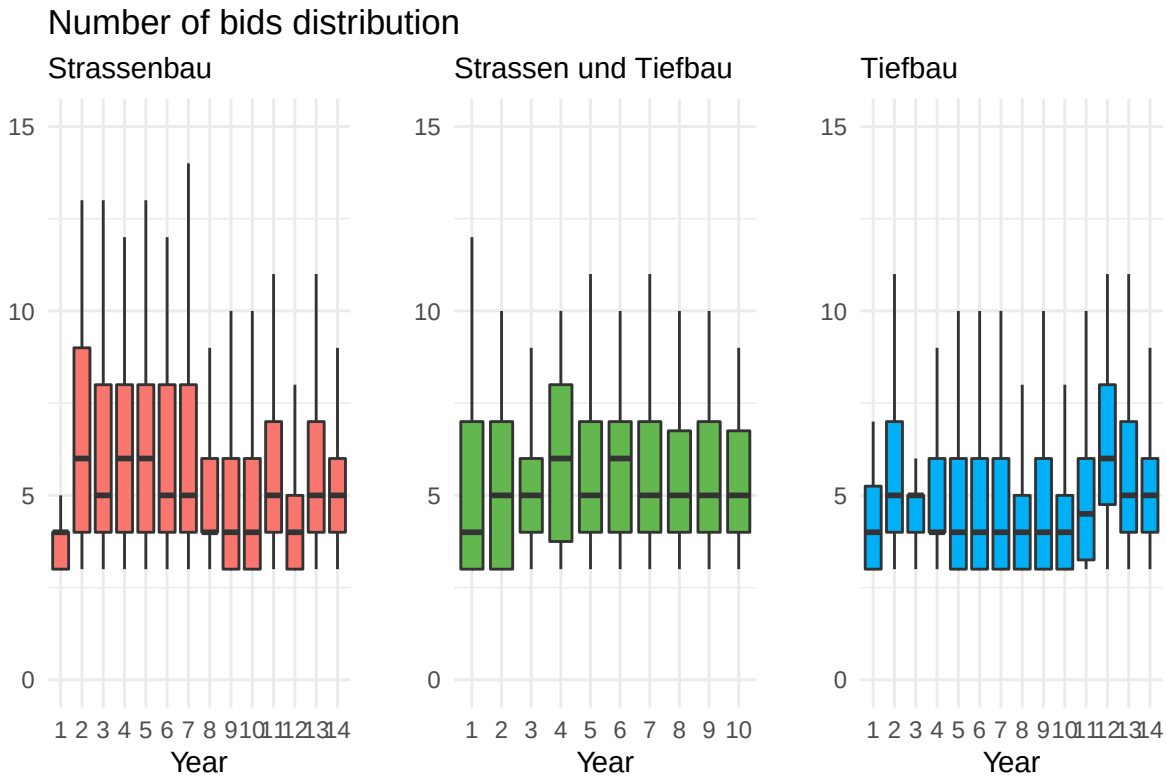
```
box1 <- box_plot(df_agg, "win_bid", 1, "#f8766d", "Strassenbau", 0, 3*10^6, " ") +
  scale_y_continuous(name="CHF", labels = scales::comma)
box2 <- box_plot(df_agg, "win_bid", 2, "#62B74E", "Strassen und Tiefbau", 0, 3*10^6," ") +
  scale_y_continuous(name="", labels = scales::comma)
box3 <- box_plot(df_agg, "win_bid", 3, "#00b0f6", "Tiefbau", 0, 3*10^6, "Note: Outliers are hidden") +
  scale_y_continuous(name="", labels = scales::comma)

box1 + labs(title = "Winning bid distribution") + box2 + box3
```



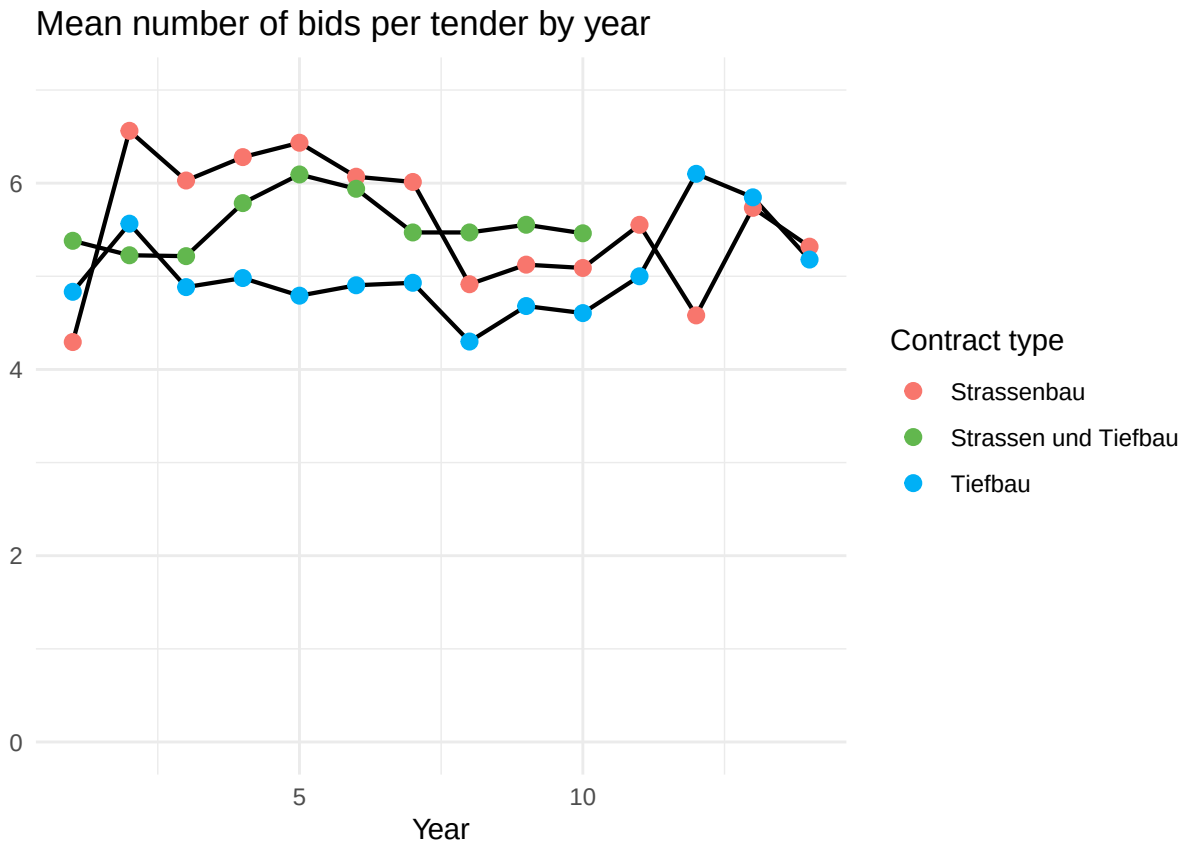

```
ggsave("images/bid_distr.png", width = 24, height = 14, units = "cm")
```

Next, how does the number of bids change over time?

```
box1 <- box_plot(df_agg, "no_bids", 1, "#f8766d", "Strassenbau", 0, 15, " ")
box2 <- box_plot(df_agg, "no_bids", 2, "#62B74E", "Strassen und Tiefbau", 0, 15," ")
box3 <- box_plot(df_agg, "no_bids", 3, "#00b0f6", "Tiefbau", 0, 15, "Note: Outliers are hidden")

box1 + labs(title = "Number of bids distribution") + box2 + box3
```




```r
ggsave("images/no_bids.png", width = 24, height = 14, units = "cm")

# Also add average number of bids:
avg_var_year(df_agg, "no_bids") +
  labs(y = "", x = "Year", title = "Mean number of bids per tender by year")  +
  scale_color_manual(name = "Contract type", values=c("#f8766d", "#62B74E", "#00b0f6"),
                     labels = c("Strassenbau", "Strassen und Tiefbau", "Tiefbau")) +
  coord_cartesian(xlim = c(1, 14), ylim = c(0, 7))
```

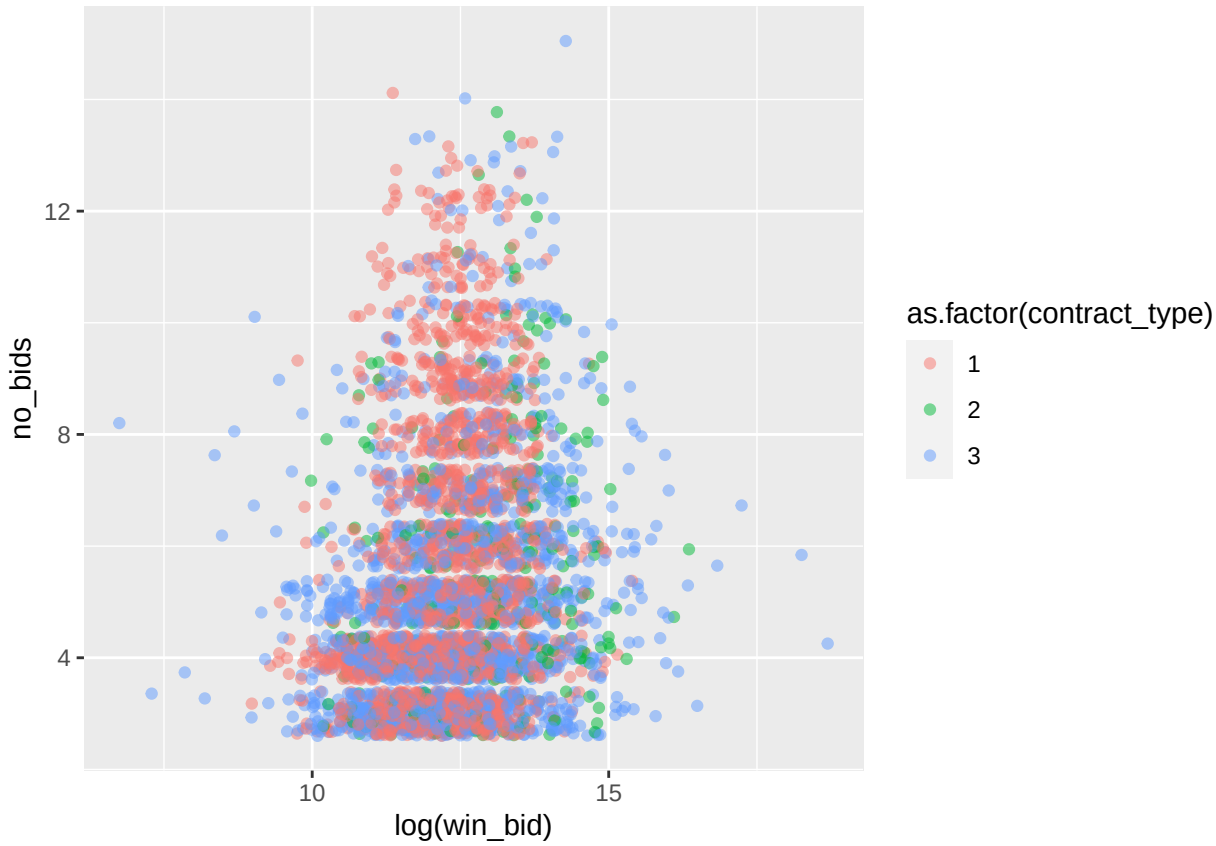



```
ggsave("images/no_bids_means.png", width = 24, height = 14, units = "cm")
```

The number of bids is fairly constant around 5 or 6 over time, but based on the boxplots, does exhibit upward skewness meaning that there for some projects are many bidders.

Can this be explained by project size?

```
df_agg %>% ggplot(aes(x = log(win_bid), y = no_bids, color = as.factor(contract_type))) +
  geom_jitter(alpha = 0.5)
```



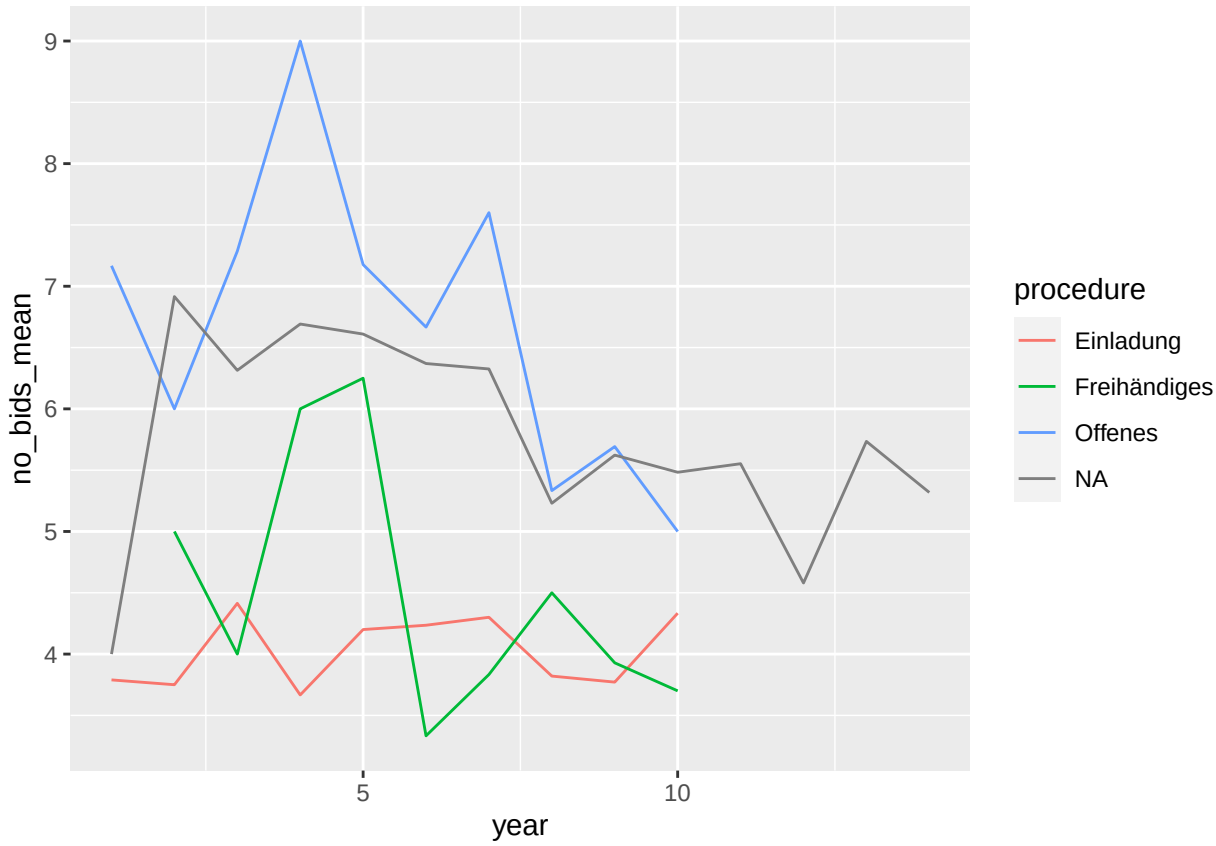

```
cor(df_agg$no_bids, log(df_agg$win_bid))
```

```
## [1] 0.1778515
```

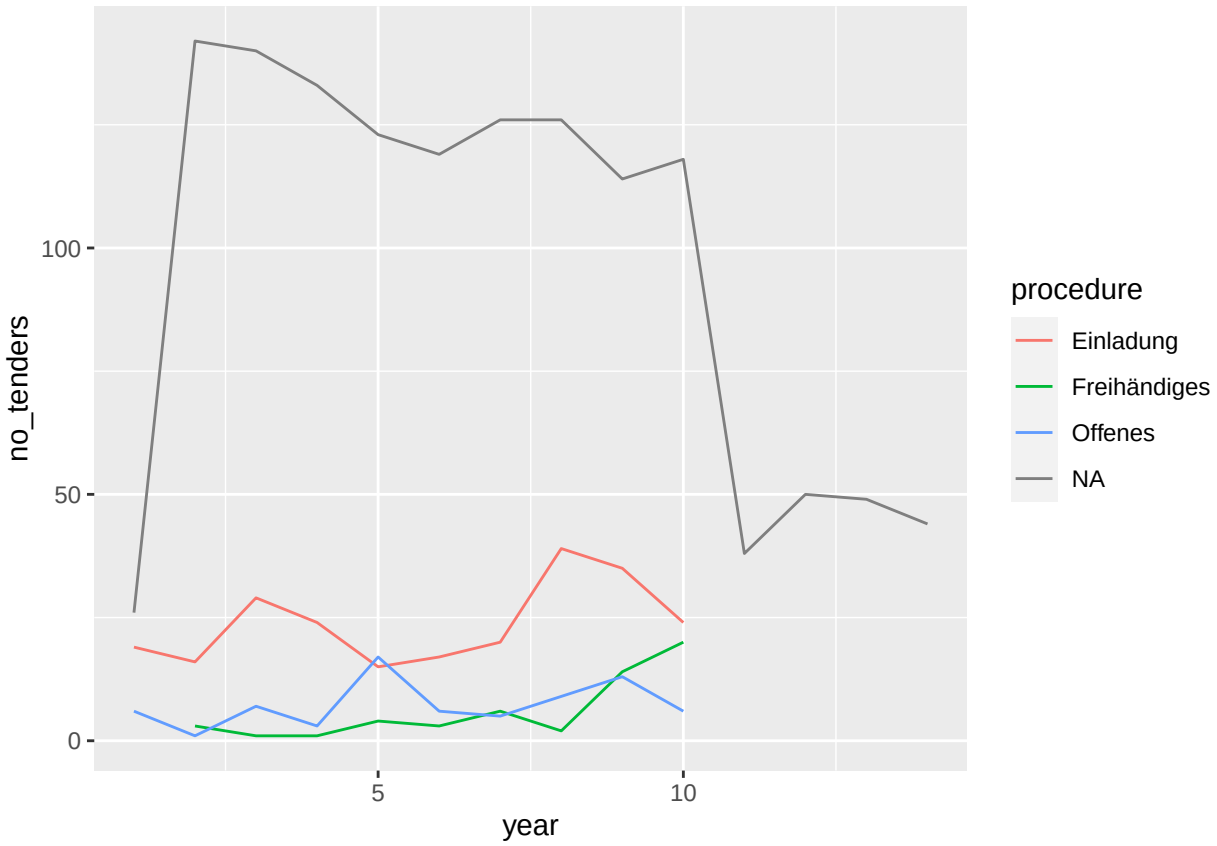
There is a small but positive correlation. Not definitely the reason why there is some variation.

How does the procedure type change over time?

```
# First, number of bids by type:
df_agg %>% filter(contract_type == 1) %>% group_by(procedure, year) %>%
  summarise(no_bids_mean = mean(no_bids)) %>%
  ggplot(aes(x = year, y = no_bids_mean, color = procedure)) +
  geom_line()
```



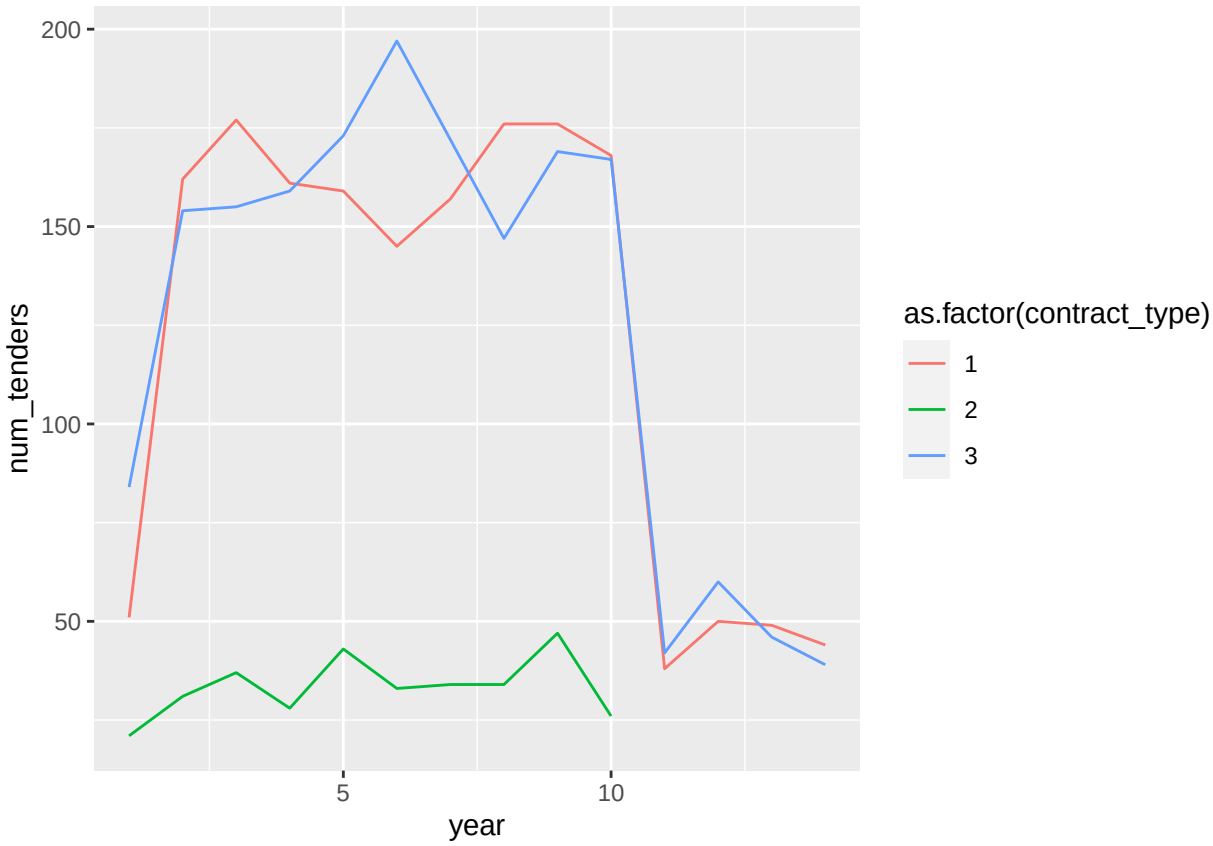

```
# Then, frequency of types:
df_agg %>% filter(contract_type == 1) %>% group_by(procedure, year) %>%
  summarise(no_tenders = n()) %>%
  ggplot(aes(x = year, y = no_tenders, color = procedure)) +
  geom_line()
```



It does not seem to be the case that the procedure type changes and explains the drop in number of bids.

Now, let's look at the total number of tenders over time:

```
df_agg_y <- df_agg %>% group_by(year, contract_type) %>% summarise(mean_value = mean(win_bid),
                                          sd_value = sd(win_bid),
                                          num_tenders = n())

df_agg_y %>% ggplot(aes(x = year, num_tenders, color = as.factor(contract_type))) +
  geom_line()
```

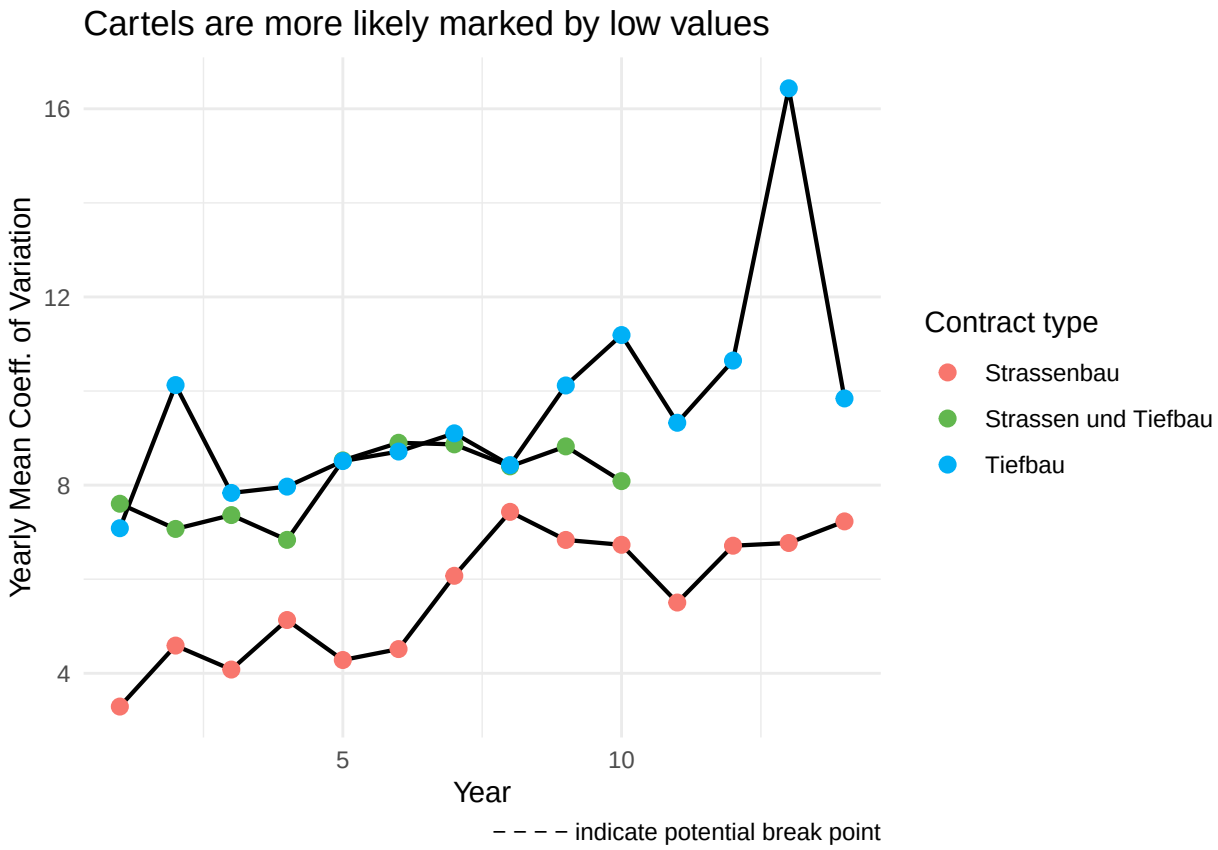


Most tenders are within the pure categories during years 1 to 10. We can expect values to be the most "stable" for the pure contract types in the early years.

First behavioural screen, time series and structural breaks

Now, we move on to focus on the time variation of bidding behaviour.

```
CV_plot <- avg_var_year(df_agg, "CV") +
  labs(y = "Yearly Mean Coeff. of Variation", x = "Year",
       title = "Cartels are more likely marked by low values",
       caption = "- - - - indicate potential break point")
CV_plot
```

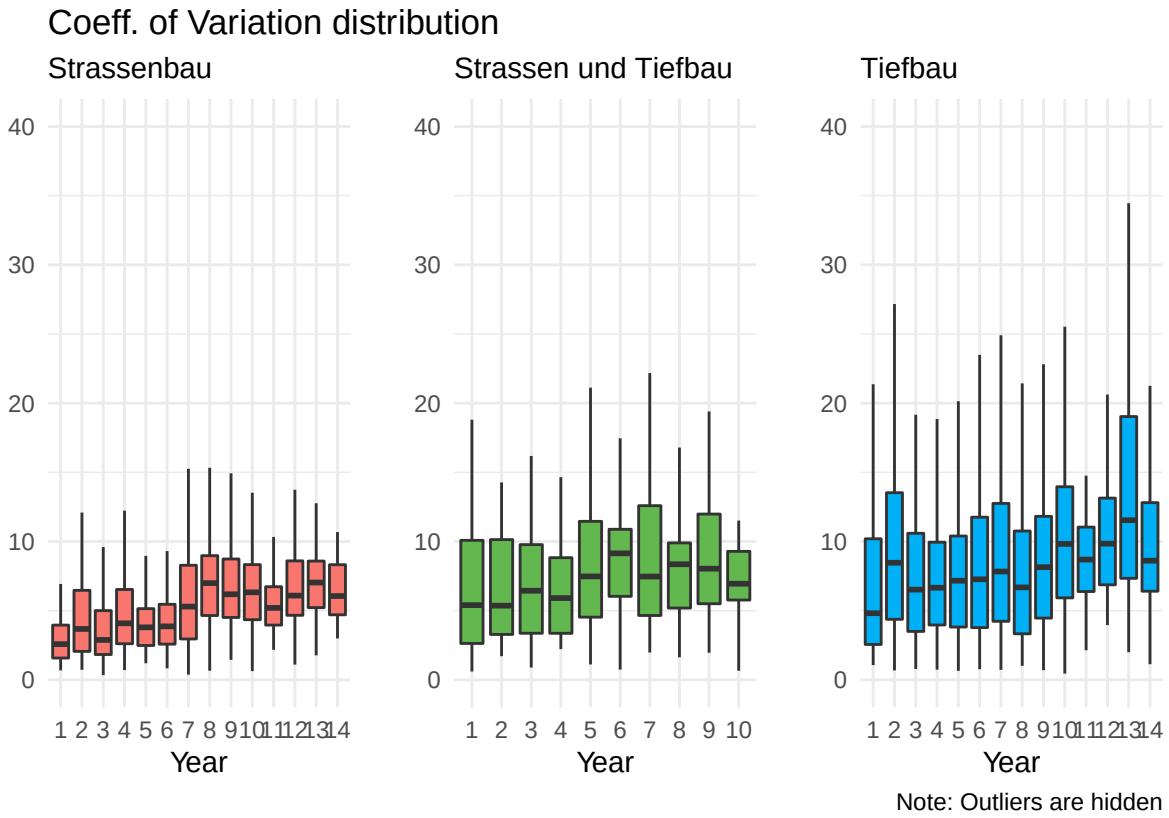



```
ggsave("images/CV_year.png", width = 24, height = 14, units = "cm")
```

The CV is the lowest for Strassenbau, suggesting this type is most prone to collusion. In addition, the level seems to be the lowest in the early years.

Dig deeper using box plots:

```
box1 <- box_plot(df_agg, "CV", 1, "#f8766d", "Strassenbau", 0, 40, " ")
box2 <- box_plot(df_agg, "CV", 2, "#62B74E", "Strassen und Tiefbau", 0, 40, " ")
box3 <- box_plot(df_agg, "CV", 3, "#00b0f6", "Tiefbau", 0, 40, "Note: Outliers are hidden")

box1 + labs(title = "Coeff. of Variation distribution") + box2 + box3
```

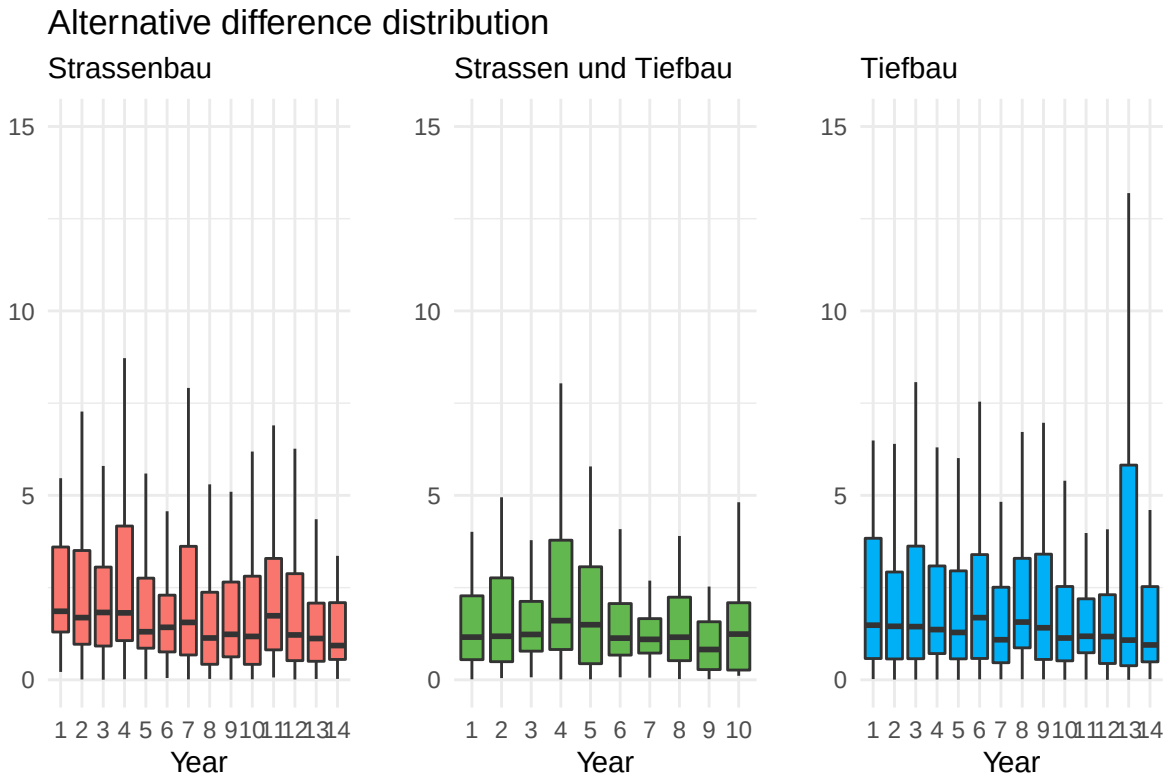



```
ggsave("images/CV_distr.png", width = 24, height = 14, units = "cm")
```

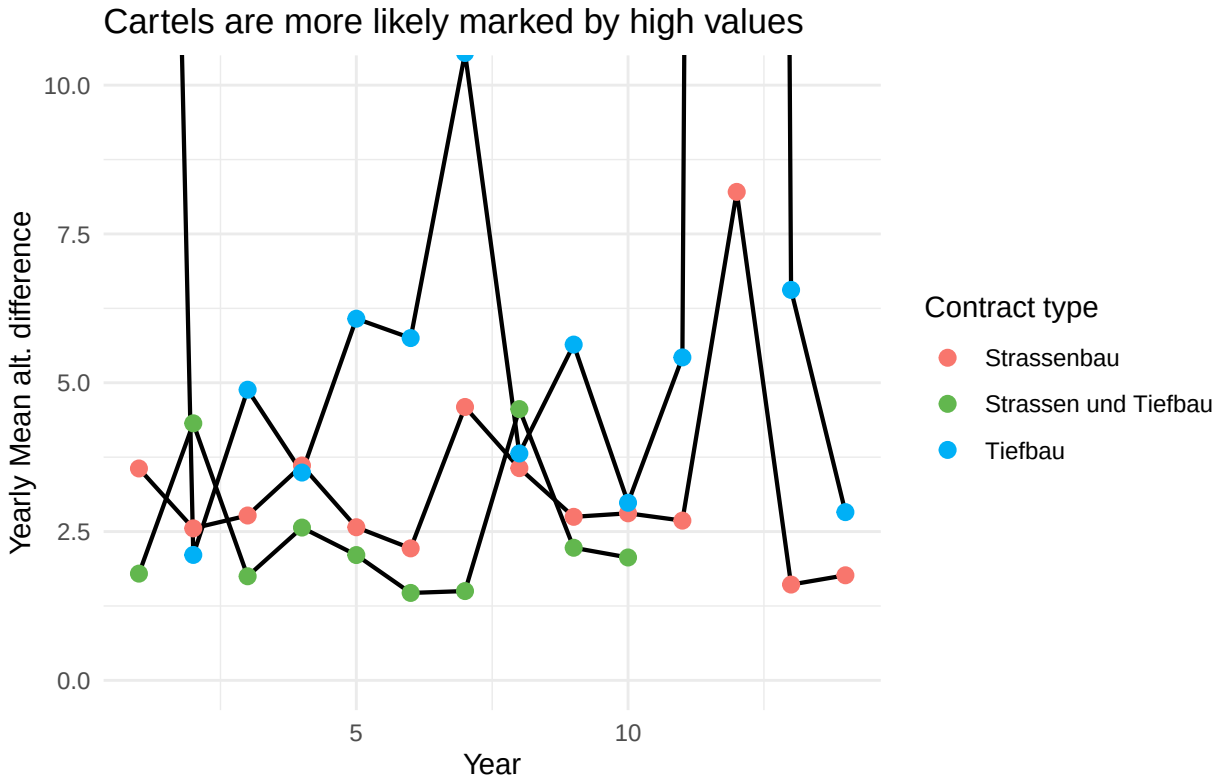
Also the median seems consistent with the mean presented above. It also looks a bit surpressed for CV in the early years compared to the other categories where values frequent around 20.

How does the altrd develop over time?

```
box1 <- box_plot(df_agg, "altrd", 1, "#f8766d", "Strassenbau", 0, 15, " ")
box2 <- box_plot(df_agg, "altrd", 2, "#62B74E", "Strassen und Tiefbau", 0, 15, " ")
box3 <- box_plot(df_agg, "altrd", 3, "#00b0f6", "Tiefbau", 0, 15, "Note: Outliers are hidden")

box1 + labs(title = "Alternative difference distribution") + box2 + box3
```



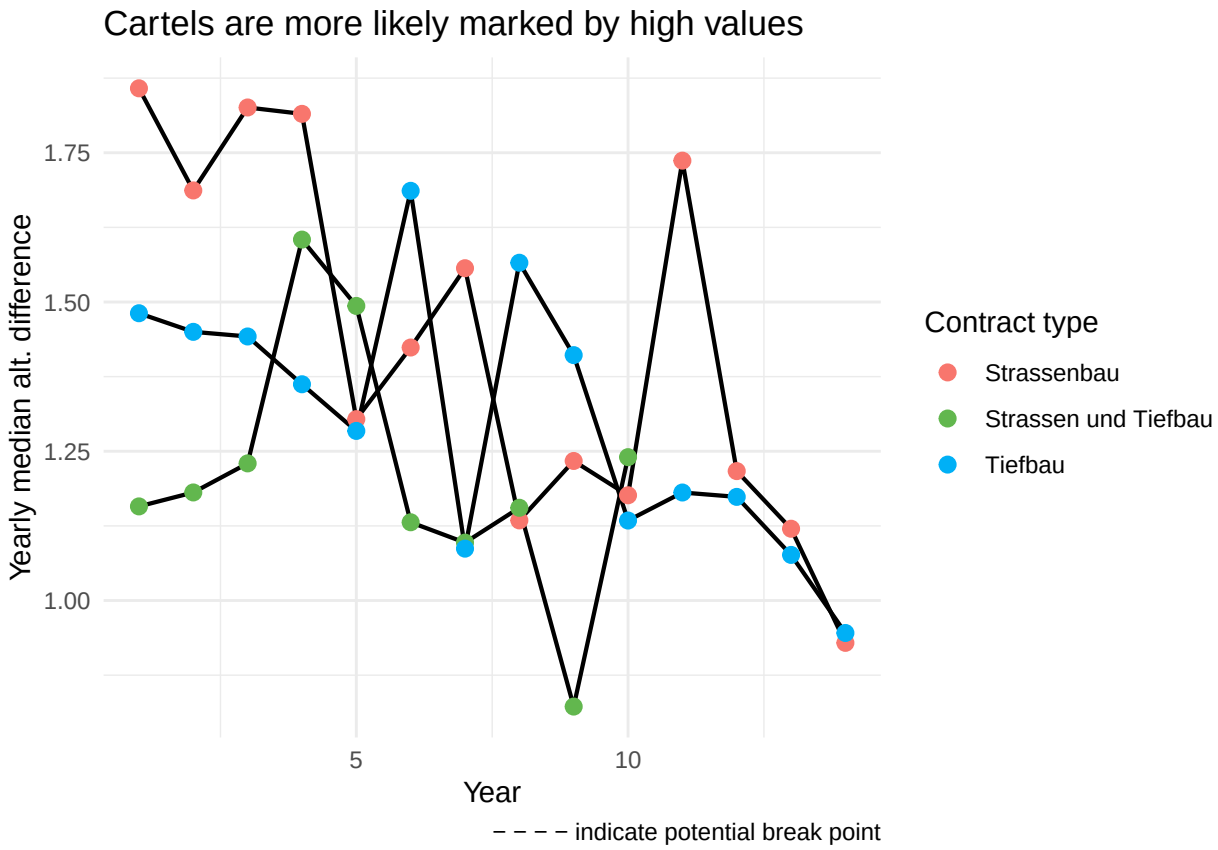

```
ggsave("images/altrd_distr.png", width = 24, height = 14, units = "cm")

altrd_plot <- avg_var_year(df_agg, "altrd") +
  labs(y = "Yearly Mean alt. difference", x = "Year",
       title = "Cartels are more likely marked by high values",
       caption = "- - - - indicate potential break point
       In years 1 and 12, Tiefbau is largely driven by two extreme outliers ") +
  coord_cartesian(ylim = c(0, 10))
altrd_plot
```




```
ggsave("images/altrd_mean.png", width = 24, height = 14, units = "cm")
```

The last graph showing means is much too noisy. We know this happens because of two outlier values as we will see later. To avoid this problem, we chose to instead look at the medians:

```
altrd_plot <- df_agg %>% select(project, year, contract_type, altrd) %>%
  group_by(year, contract_type) %>%
  summarise(altrd_median = median(altrd)) %>%
  ggplot(aes(x = year, y = altrd_median)) +
  geom_line(size = 0.75, aes(fill = as.factor(contract_type))) +
  geom_point(size = 2.5, aes(color=factor(contract_type))) +
  theme_minimal() +
  scale_color_manual(name = "Contract type", values=c("#f8766d", "#62B74E", "#00b0f6"),
                     labels = c("Strassenbau", "Strassen und Tiefbau", "Tiefbau")) +
  labs(y = "Yearly median alt. difference", x = "Year",
       title = "Cartels are more likely marked by high values",
       caption = "- - - - indicate potential break point")
altrd_plot
```

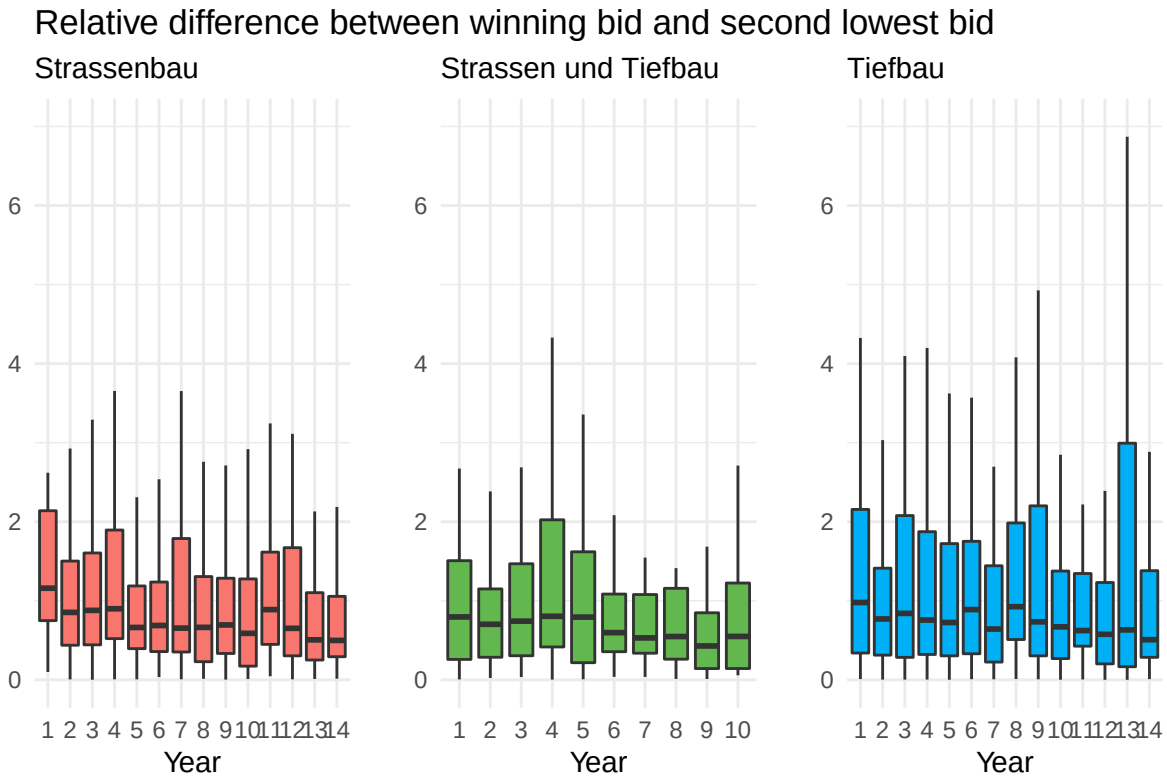



```
ggsave("images/altrd_median.png", width = 24, height = 14, units = "cm")
```

This plot is also very noisy unfortunately, but at least everything is scaled well. Gives a sense that altrd is decreasing over time which would point towards cartels being more prevalent in the early years. Again, Strassenbau stands out as the highest and thereby the one that looks most collusive.

Is rd different?

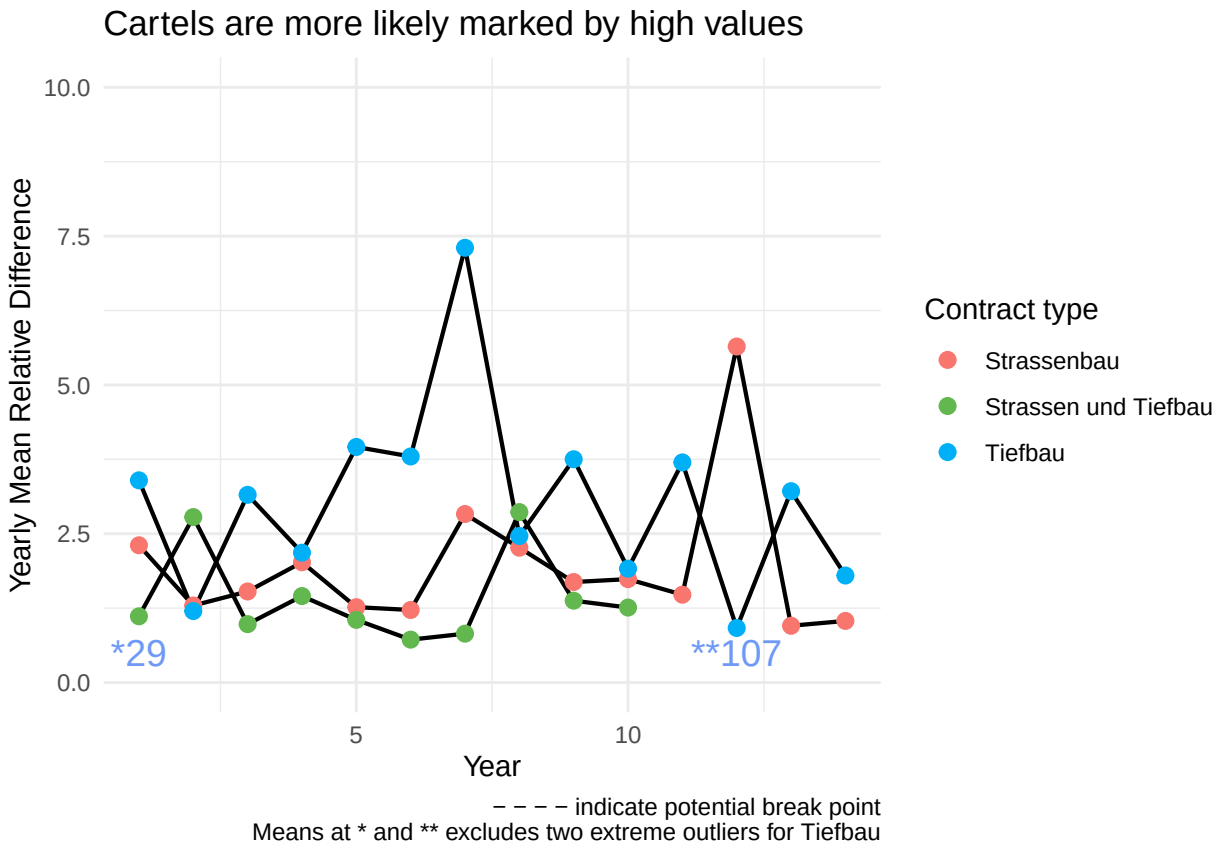
```
box1 <- box_plot(df_agg, "rd", 1, "#f8766d", "Strassenbau", 0, 7, " ")
box2 <- box_plot(df_agg, "rd", 2, "#62B74E", "Strassen und Tiefbau", 0, 7, " ")
box3 <- box_plot(df_agg, "rd", 3, "#00b0f6", "Tiefbau", 0, 7, "Note: Outliers are hidden")

box1 + labs(title = "Relative difference between winning bid and second lowest bid") + box2 + box3
```




```
ggsave("images/rd_distr.png", width = 24, height = 14, units = "cm")
```

```
# For rd, filter away two extreme values:
df_agg_filtered <- df_agg %>% filter(rd < 1000)
 #Actual values:
rd_12 <- df_agg %>% select(project, year, contract_type, rd) %>%
  filter(contract_type == 3, year == 12) %>%
  group_by(year) %>%
  summarise(avg_rd = mean(rd)) %>% select(avg_rd) %>% as.integer()

rd_1 <- df_agg %>% select(project, year, contract_type, rd) %>%
  filter(contract_type == 3, year == 1) %>%
  group_by(year) %>%
  summarise(avg_rd = mean(rd)) %>% select(avg_rd) %>% as.integer()


annotation <- data.frame(
   x = c(1,12),
   y = c(0.5,0.5),
   label = c(glue::glue("*", round(rd_1)), glue::glue("**", round(rd_12)))
)

rd_plot <- avg_var_year(df_agg_filtered, "rd") +
  labs(y = "Yearly Mean Relative Difference", x = "Year",
       title = "Cartels are more likely marked by high values",
       caption = "- - - - indicate potential break point
Means at * and ** excludes two extreme outliers for Tiefbau") +
    coord_cartesian(ylim = c(0, 10)) +
```

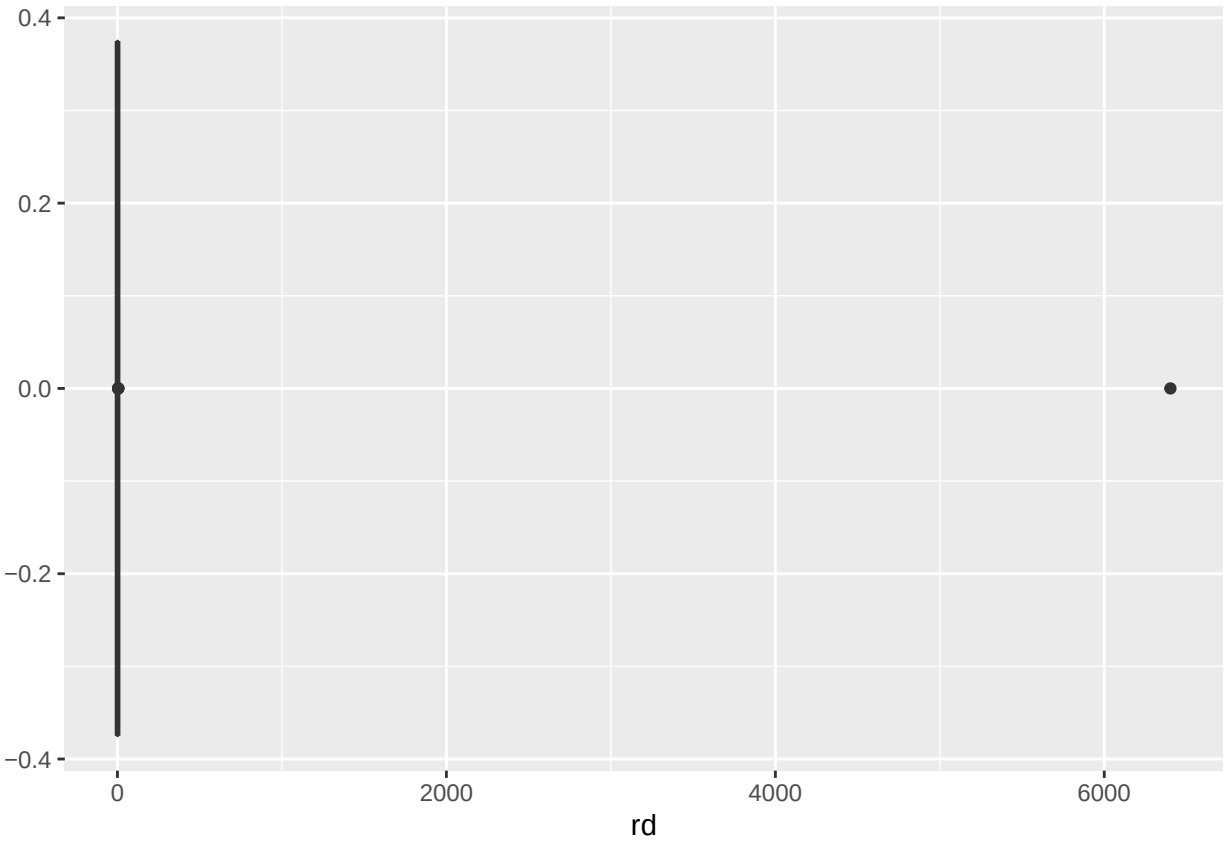
```
  geom_text(data=annotation, aes(x=x, y=y, label=label), color="#6e99f7",
            size=5)

rd_plot
```




```
ggsave("images/rd_year.png")
```

Why is RD for Tiefbau so high at year 12?

```
df_agg %>% filter(contract_type == 3, year == 12) %>% select(rd) %>%
  ggplot(aes(x = rd)) +geom_boxplot()
```



One value seem to be circa 6000

```
df_agg %>%
  filter(contract_type == 3, year == 12) %>% select(rd, project) %>% arrange(-rd)
```

```
## # A tibble: 60 x 2
##         rd project
##      <dbl>   <dbl>
##  1 6402.      4154
##  2    5.34    4201
##  3    4.63    4178
##  4    4.43    4148
##  5    3.71    4106
##  6    2.39    4184
##  7    2.00    4187
##  8    1.74    4215
##  9    1.72    4169
## 10    1.61    4189
## # ... with 50 more rows
```

```
# This is project with id 4154
df %>% filter(project == 4154)
```

```
##        bid contract_type procedure project anonymisedyear anonymiseddate
## 1 162818.4             3      <NA>    4154             12           4136
## 2 177305.2             3      <NA>    4154             12           4136
## 3 177308.5             3      <NA>    4154             12           4136
```

Apparently, the variance of the loosing bids is extremely low. This also explains why altrd was so extreme

when we looked at the means.

Second behavioural screen, fit model

In this screen, we apply a pre developed model by Huber and Imhof, "Machine Learning with Screens for Detecting Bid-Rigging Cartels" (2018)

Apply model

In the paper, the authors fit a logistic regression and give the coefficients. We simply use these coefficients to generate predictions. Important, this implicitly assumes our data come from the same distribution as the data found in the ML paper. This assuption must be discussed when we present our evidence.

Also, note that this is only an exploratory anb rough draft so far and that there are some things to do before the actual graphs can be produced.

```
# Alternatice way: force coefficients: https://tolstoy.newcastle.edu.au/R/e2/help/07/08/24294.html
m3 <- c(1.02, -0.49, 0.92, 0.09)
m4 <- c(1.51, -0.47, 0.95, 0)
mlasso <- c(0.82, -0.394, 0.699, 0.039)
names(m3) <- c("CONST", "CV", "ALTRD", "NoBIDS")
names(m4) <- names(m3)
names(mlasso) <- c("CONST", "CV", "ALTRD", "NoBIDS")
# Use model 4 because number of bids is insignificant in model 3
# While lasso would correspond best with the confusion matrix, it cannot be used:
# First, lambda is not reported, second, the constant is not reported
# (we used the one from model 1)

model <- m4

df_agg %>%
  mutate(
    probability_collusion = 1 / (1 + exp(-(model["CONST"] + model["CV"] * CV +
                                             model["ALTRD"] * altrd +
                                             model["NoBIDS"] * no_bids)))
    ) -> df_agg
```

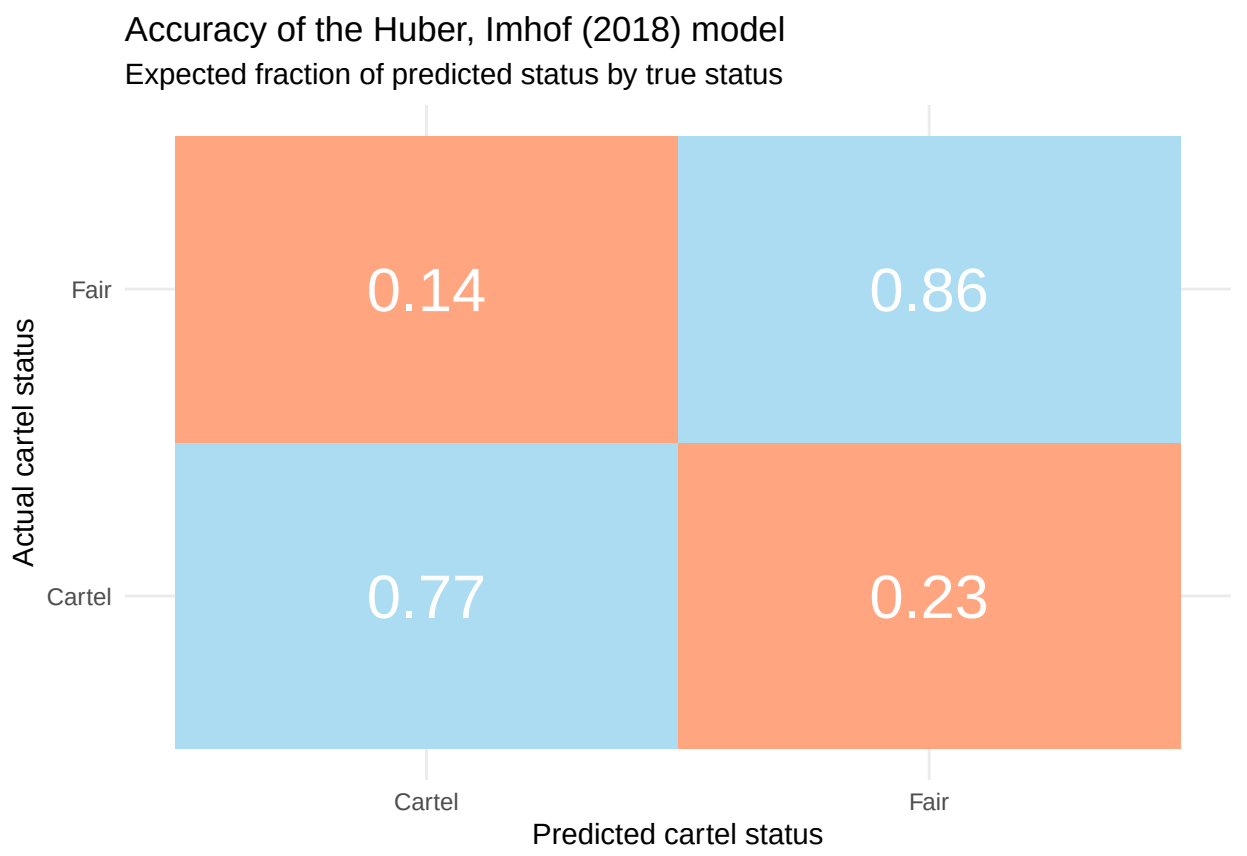
Create confusion matrix

Based on Figure 3 in the paper, we can get a sense of how good the classifier is. Unfortunately, it does not report results for the model where we have the coefficients available. Anyhow, it is a usual conceptual tool and the two models (the logisitc regression we use, and the lasso logistic regression on which the confusion matrix is based) share the same variables and functional form and are thus expected to yield similar results.

```
# Data from ML paper fig 3
# Note that it works for different decision criteria
TP <- c(1, 0.91, 0.85, 0.77, 0.64, 0.38)
TN <- c(0, 0.69, 0.78, 0.86, 0.91, 0.97)
FP <- 1 - TN
FN <- 1 - TP
dc_vec <- c(0, 0.5, 0.6, 0.7, 0.8, 0.9)
certainty_info <- cbind(dc_vec, TP, TN, FP, FN) %>% as_tibble()
```

```r
# Use decision criteria, dc, to classify observations
df_agg <- df_agg %>% mutate(pred_cartel = if_else(probability_collusion > dc, 1, 0))
```

```r
dc_ix <- which(dc_vec == dc)
num_cartel_pred <- df_agg %>% filter(probability_collusion >= dc) %>% nrow()
num_fair_pred <- df_agg %>% filter(probability_collusion < dc) %>% nrow()

# Prepare data
conf_actual <- c("Cartel", "Fair")
conf_pred <- c("Cartel", "Fair")
conf_mat <- expand.grid(X=conf_pred, Y=conf_actual)
conf_mat$conf_data <- c(TP[[dc_ix]] , FN[[dc_ix]],
                        FP[[dc_ix]], TN[[dc_ix]])
conf_mat$colors <- c(0,1,1,0)

# Plot the confusion matrix
conf_mat %>% ggplot(aes(conf_mat$X, conf_mat$Y)) +
  geom_tile(aes(fill = colors)) +
  labs(title =  "Accuracy of the Huber, Imhof (2018) model", subtitle = "Expected fraction of predicted
       y = "Actual cartel status", x = "Predicted cartel status") +
  theme_minimal() +
  geom_text(aes(label = round(conf_data, 2)),size = 8, color = "white") +
  scale_fill_gradient2(low = "skyblue", high = "coral", mid = "white",
   midpoint = 0.5, limit = c(0,1.2), space = "Lab",
    name="") +
  theme(legend.position = "none")
```



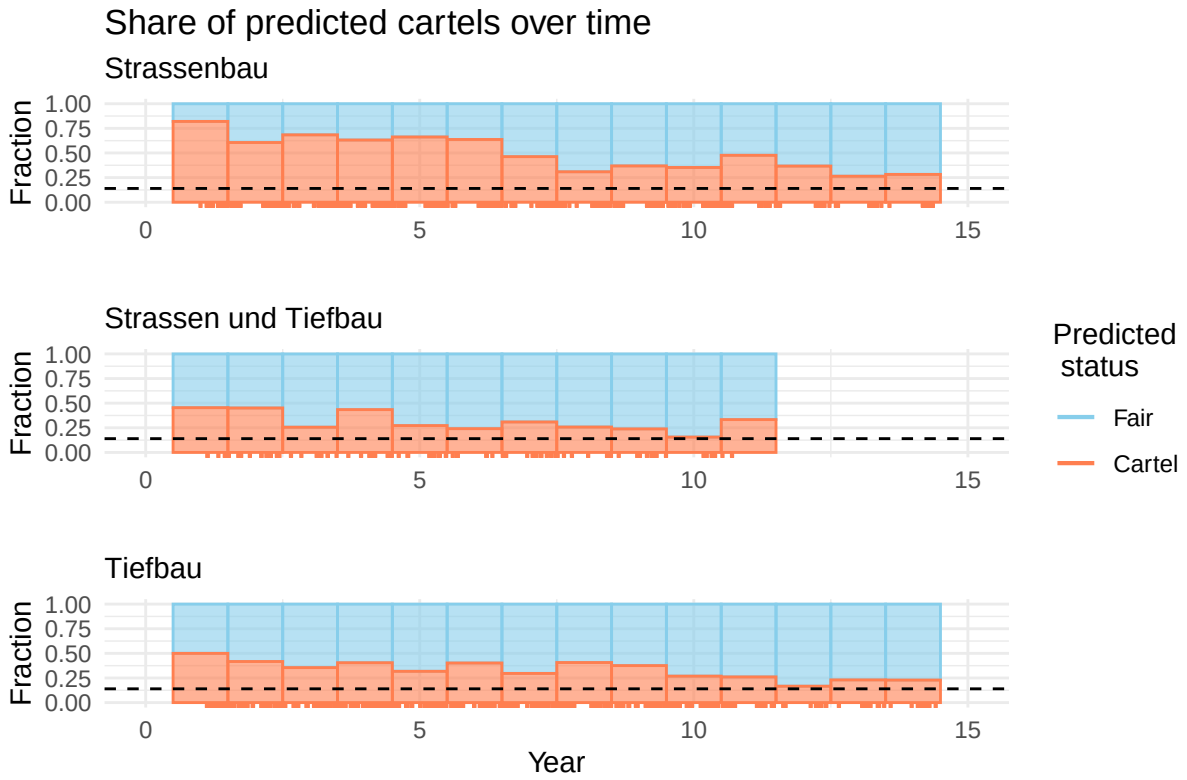
```
ggsave("images/confmat.png")
```

The confusion matrix shows us that given that a tender is cartelised, there is a 77 percent probability of labelling it as such. Whereas it labels 14 percent of those that were fair as cartels. This applies if we believe the distribution is the same as in the original paper and that the LASSO and logit models are of similar accuracy.

Plot densities of occurences.

Here, we look at how the density is spread over time. Idea is to capture periods with many collusive tenders so that we can form an opinion of whether the period was truly cartelised.

```
df_classified <- df_agg %>%
  mutate(pos = if_else(probability_collusion > dc, 1, 0),
         contr_type = contract_type,
         pos2 = pos) %>%
  unite(col = "dens_cat", c(pos2, contr_type ))

plot_dist <- function(data, contract, category_name){
  data %>% filter(contract_type %in% contract) %>%
  ggplot(aes(x = date2, color = as.factor(pos), fill = as.factor(pos))) +
  geom_histogram(alpha = 0.6, show.legend = FALSE,  binwidth = 1, position = "fill") +
    geom_rug(data = data %>% filter(pos == 1, contract_type == contract), size = 0.75) +
    scale_color_manual(name = "Predicted \n status", values=c("skyblue", "coral"), labels = c("Fair", "
    scale_fill_manual(values=c("skyblue", "coral")) +
    theme_minimal() + labs(x = "") +
    #theme(legend.position = "None") +
    labs(subtitle = category_name, y = "Fraction") +
    geom_hline(yintercept=FP[[dc_ix]], linetype="dashed",
                color = "black", size=0.5) +
    xlim(min(floor(data$date2)) - 1, max(floor(data$date2)) +1)
}

dist_cat1 <- plot_dist(df_classified, 1, "Strassenbau") + theme(legend.position = "None") + labs(title =
dist_cat2 <- plot_dist(df_classified, 2, "Strassen und Tiefbau")
dist_cat3 <- plot_dist(df_classified, 3, "Tiefbau") + theme(legend.position = "None")

dist_cat1 / dist_cat2 / dist_cat3 +
  labs(caption = "Note: - - - - denotes the fraction of expected predicted cartels given no cartel (fals
```



Note: – – – – denotes the fraction of expected predicted cartels given no cartel (false positives)

```
ggsave("images/densities.png")
```

This graph confirms what we previously saw. Namely that Strassenbau is the most suspicious category and that it is in particular the early years we should be concerned about. The 14 percent benchmark looks low, as there are two marked periods in Strassenbau, a high and a low, where the low is close to 30 percent or more. While this could be explained by there being a cartel that breaks down only partially, it is probably the case that the true false positive rate is much higher than 14 percent. This implies that Tiefbau is not collusive despite consistenly being above the presumed 14 percent line.

Next, we test whether the distribution of cartelised bids can be said to be uniform. If not, it is evidence that there is a structural break somewhere.

```
kstest_contract <- function(df, type){
  # Tests whether the distribution of all predicted cartels follow the same
  # distribution as all predicted non cartels
  x <- df %>% filter(pos == 1) %>% filter(contract_type == type) %>% select(date)
  x <- x$date
  ks.test(x, "punif", min(x),max(x)) # make no use of extreme value theory here, too complicated!
}

# Test whether events deemed fair respecitve collusive occur at different points in time
kstest_contract(df_classified, 1) # Significant
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  x
## D = 0.22564, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

```
kstest_contract(df_classified, 2) # Not significant
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  x
## D = 0.090043, p-value = 0.4112
## alternative hypothesis: two-sided
```

```
kstest_contract(df_classified, 3) # Significant
```

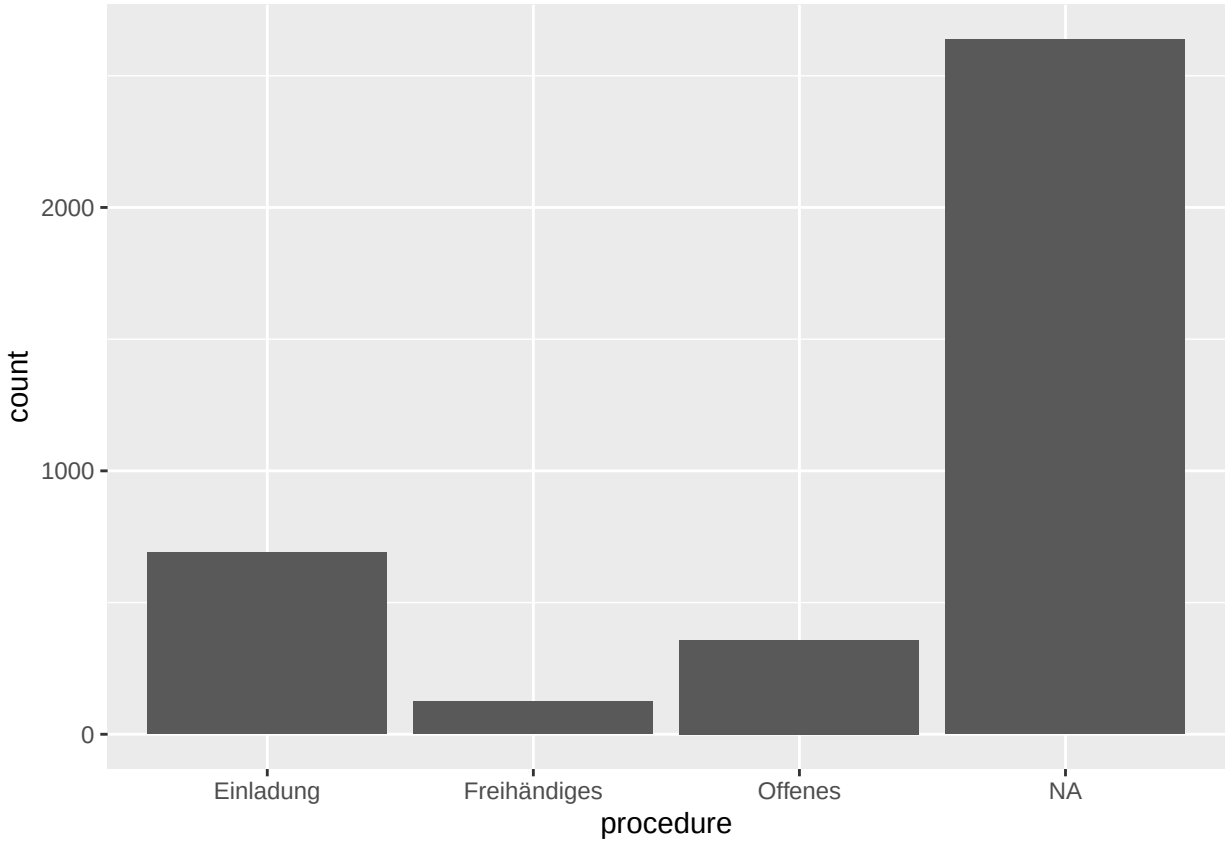
```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  x
## D = 0.21752, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

Two of the tests were significant, those for Strassenbau and Tiefbau. Visibly, Strassenbau has a structural break whereas Tiefbau sees a gradual decline.

Procedure type analysis

Here, we check for effects related to procedure type.

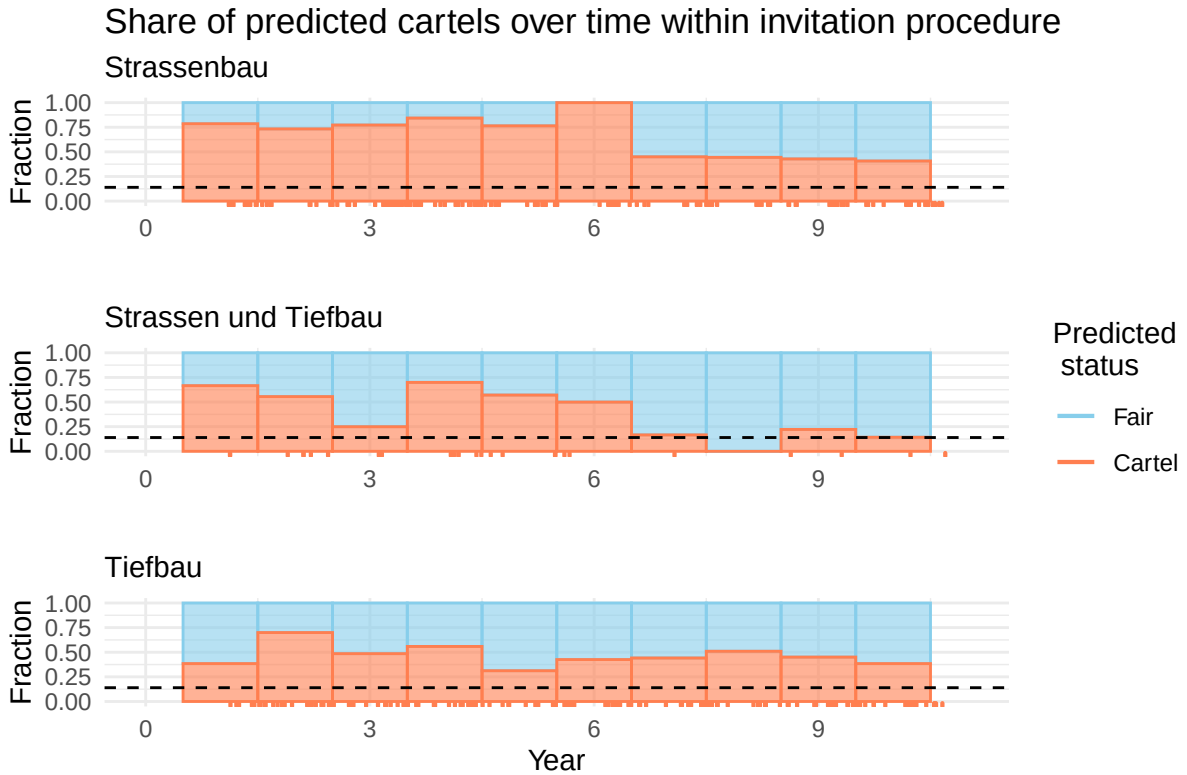
```
# What is the distribution of the different procedures?
df_agg %>% ggplot(aes(x = procedure)) + geom_bar() # Many NAs!
```



```
df_agg %>% group_by(procedure) %>% summarise(count = n())
```

```
## # A tibble: 4 x 2
##   procedure    count
##   <chr>        <int>
## 1 Einladung      690
## 2 Freihändiges   125
## 3 Offenes        358
## 4 <NA>          2638
```

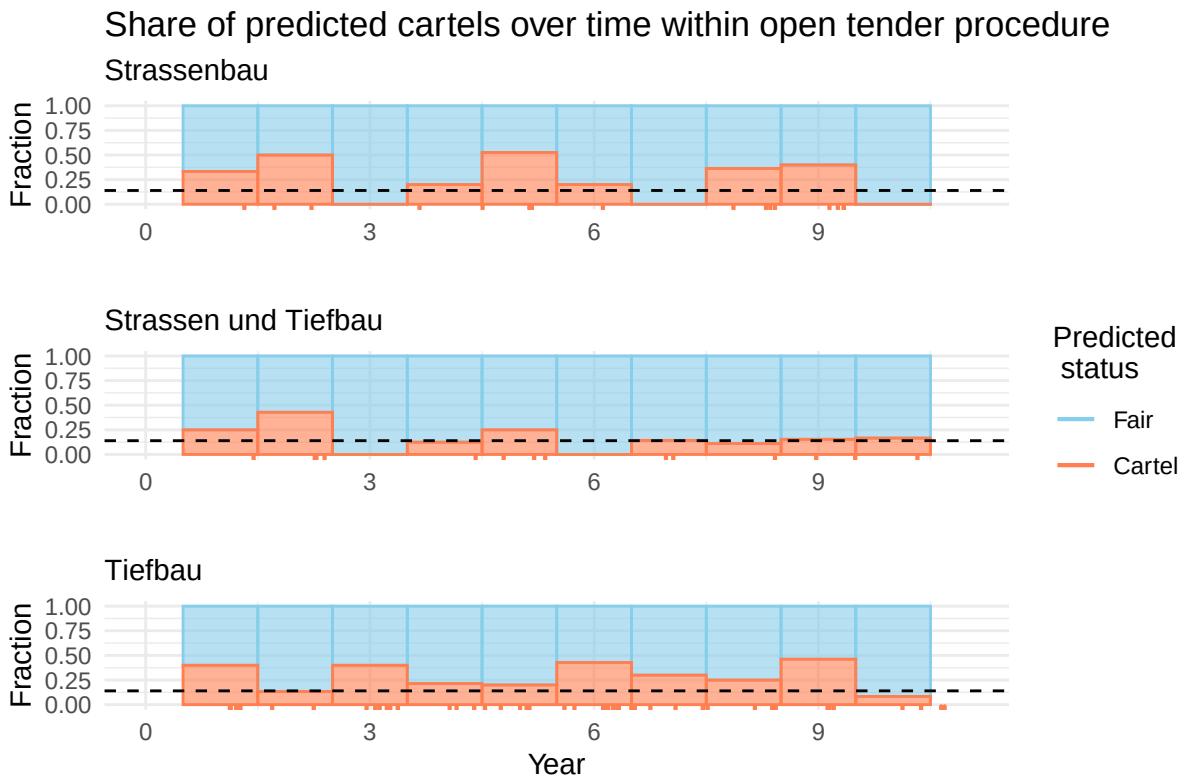
```
# Create dfs to analyse the procedures
df_einladung <- df_classified %>% filter(procedure == "Einladung")
df_offenes <- df_classified %>% filter(procedure == "Offenes")
df_other <- df_classified %>% filter(!procedure %in% c("Einladung", "Offenes"))
df_frei <- df_classified %>% filter(procedure == "Freihändiges")

# Einladung
dist_cat1 <- plot_dist(df_einladung, 1, "Strassenbau") + theme(legend.position = "None") + labs(title =
dist_cat2 <- plot_dist(df_einladung, 2, "Strassen und Tiefbau")
dist_cat3 <- plot_dist(df_einladung, 3, "Tiefbau") + theme(legend.position = "None")

dist_cat1 / dist_cat2 / dist_cat3 +
  labs(caption = "Note: - - - - denotes the fraction of expected predicted cartels given no cartel (fals
```



Note: – – – – denotes the fraction of expected predicted cartels given no cartel (false positives)

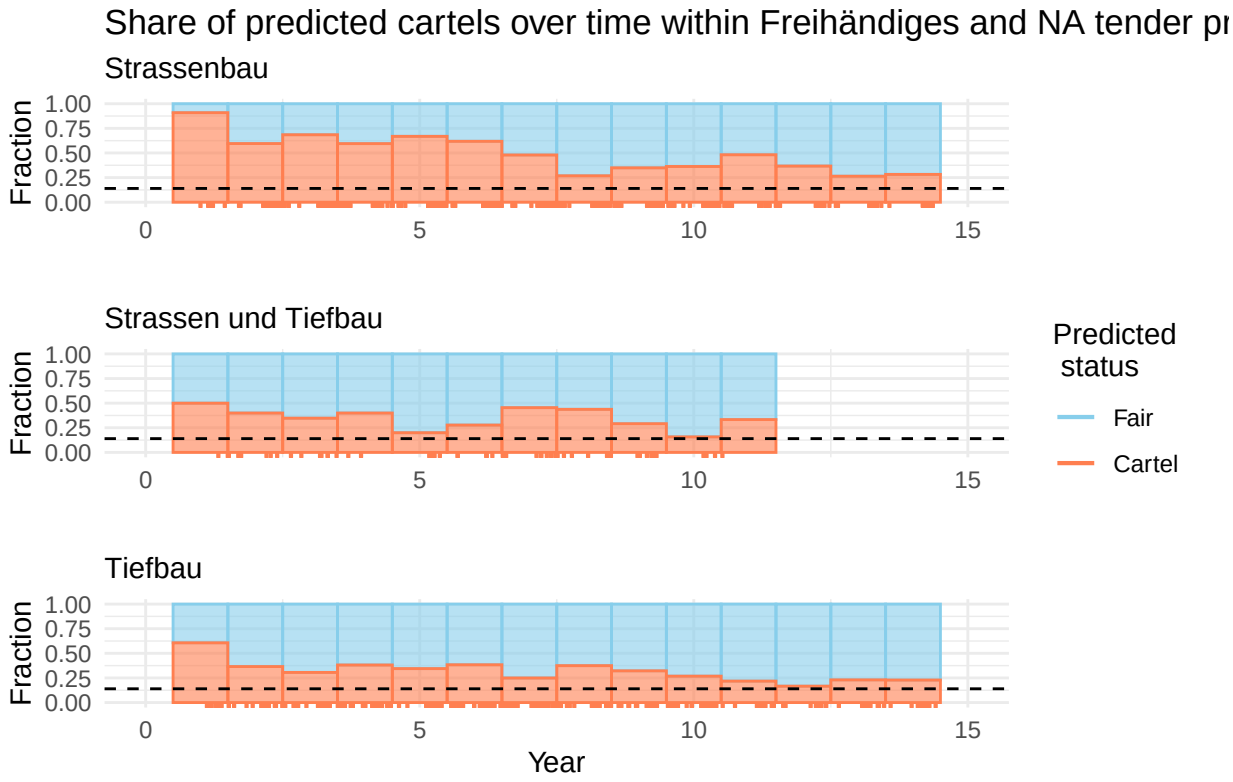
```
ggsave("images/sub/densities_ein.png")

# Offenes
dist_cat1 <- plot_dist(df_offenes, 1, "Strassenbau") + theme(legend.position = "None") + labs(title = "
```

```
dist_cat2 <- plot_dist(df_offenes, 2, "Strassen und Tiefbau")
dist_cat3 <- plot_dist(df_offenes, 3, "Tiefbau") + theme(legend.position = "None")

dist_cat1 / dist_cat2 / dist_cat3 +
  labs(caption = "Note: - - - - denotes the fraction of expected predicted cartels given no cartel (fals
```



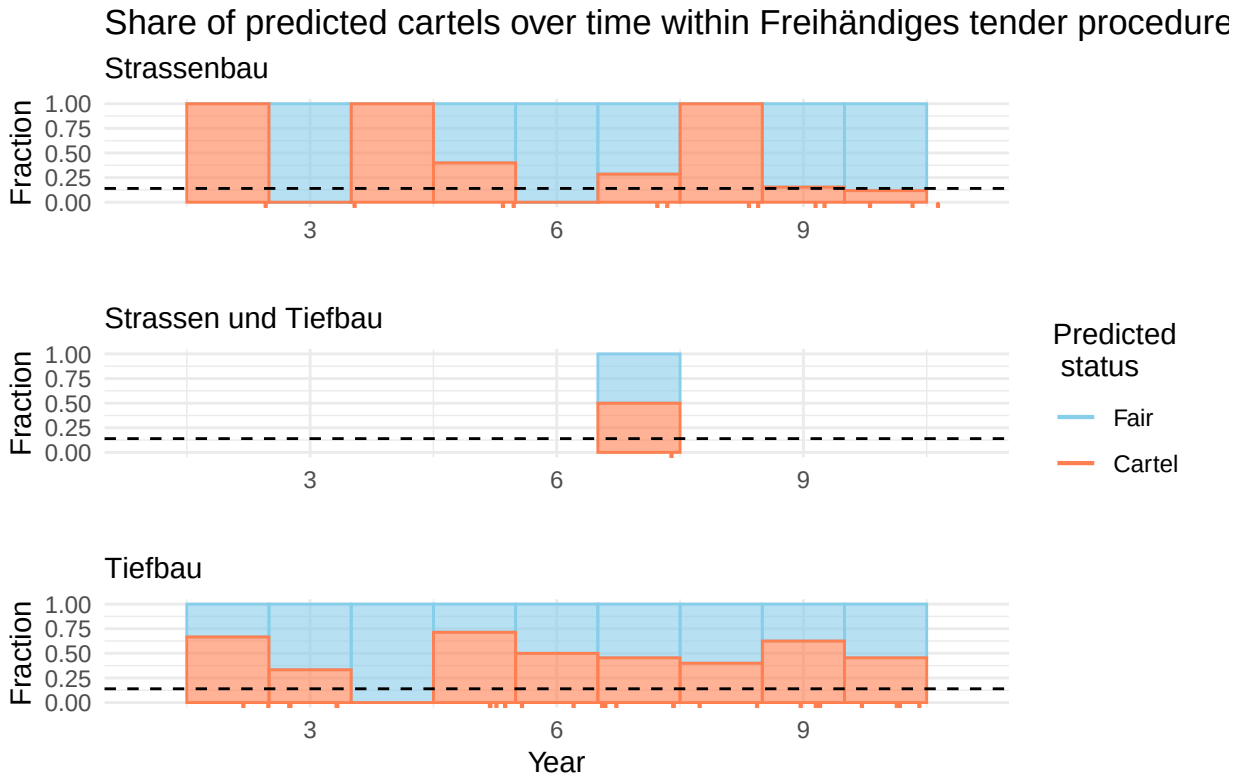
Note: – – – – denotes the fraction of expected predicted cartels given no cartel (false positives)

```
ggsave("images/sub/densities_off.png")

# Other
dist_cat1 <- plot_dist(df_other, 1, "Strassenbau") + theme(legend.position = "None") + labs(title = "Sha
dist_cat2 <- plot_dist(df_other, 2, "Strassen und Tiefbau")
dist_cat3 <- plot_dist(df_other, 3, "Tiefbau") + theme(legend.position = "None")

dist_cat1 / dist_cat2 / dist_cat3 +
  labs(caption = "Note: - - - - denotes the fraction of expected predicted cartels given no cartel (fals
```



Note: – – – – denotes the fraction of expected predicted cartels given no cartel (false positives)

```
ggsave("images/sub/densities_other.png")

# Freihändiges
dist_cat1 <- plot_dist(df_frei, 1, "Strassenbau") + theme(legend.position = "None") + labs(title = "Shar
dist_cat2 <- plot_dist(df_frei, 2, "Strassen und Tiefbau")
dist_cat3 <- plot_dist(df_frei, 3, "Tiefbau") + theme(legend.position = "None")

dist_cat1 / dist_cat2 / dist_cat3 +
  labs(caption = "Note: - - - - denotes the fraction of expected predicted cartels given no cartel (fals
```

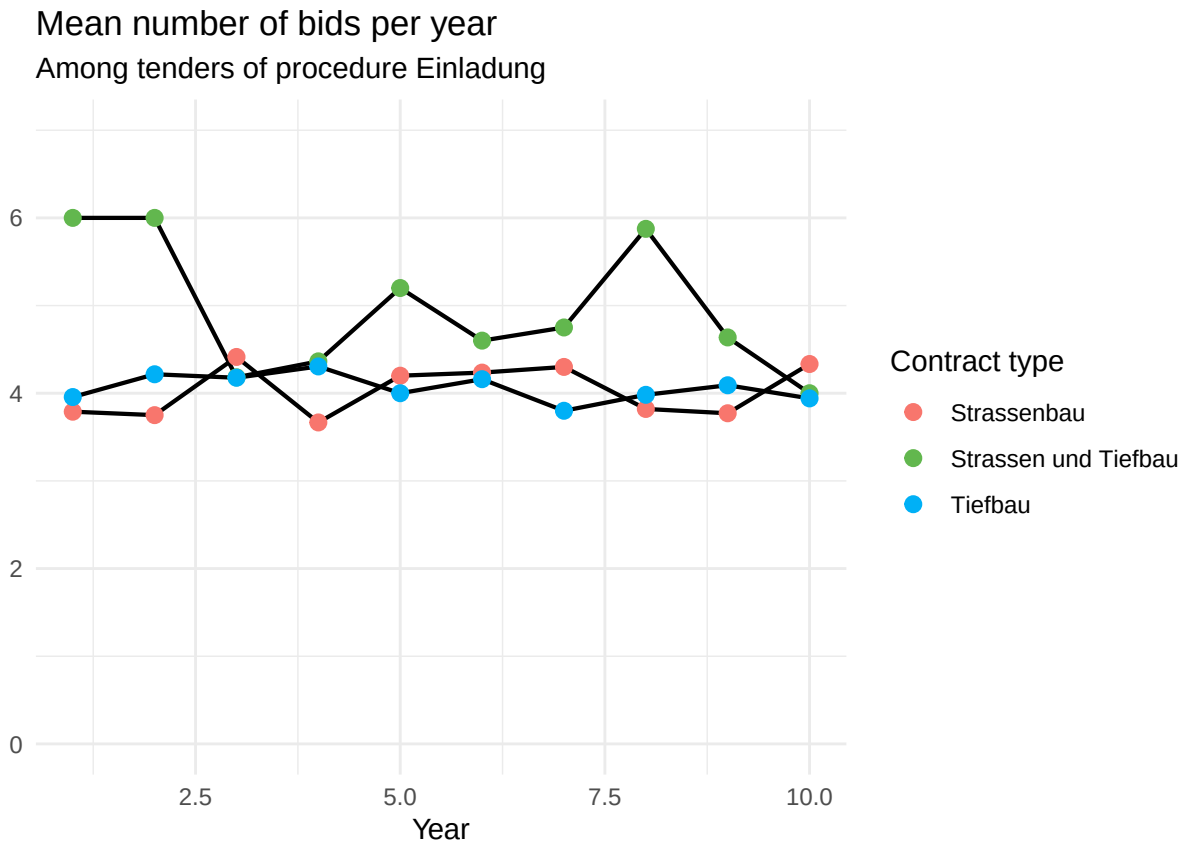


Note: – – – – denotes the fraction of expected predicted cartels given no cartel (false positives)

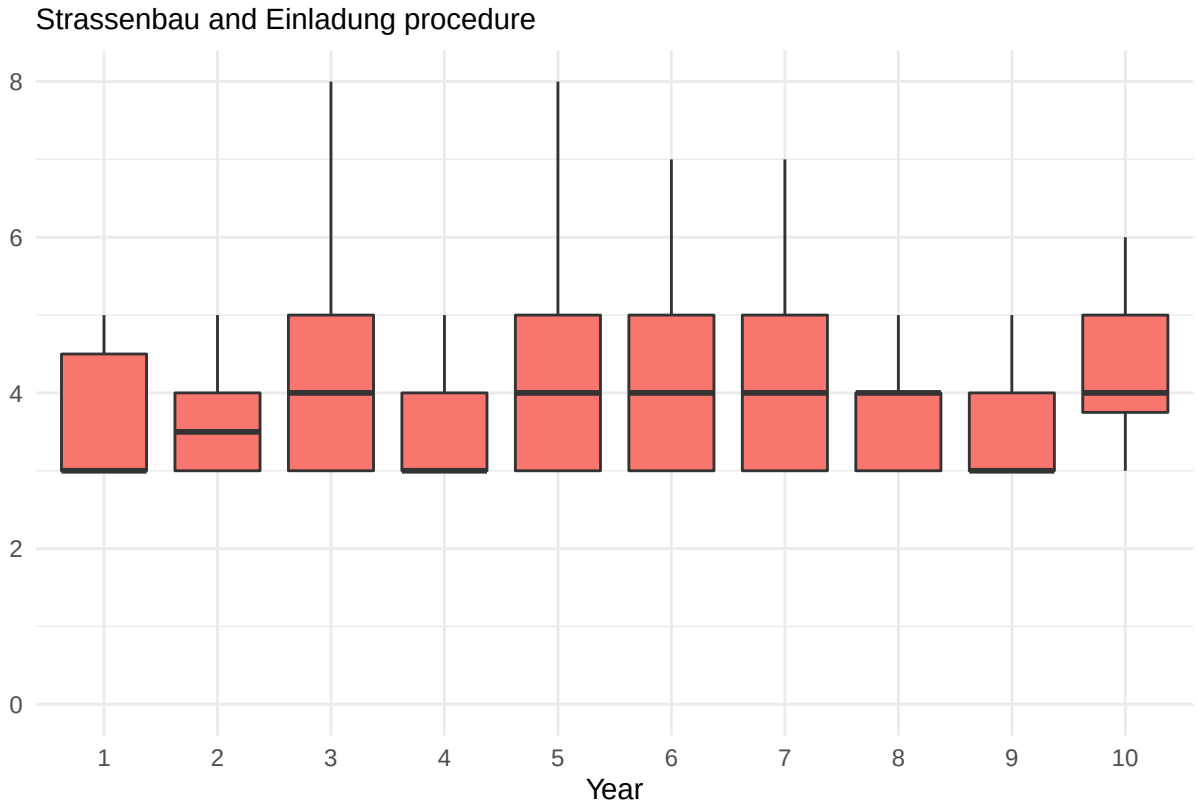
```
ggsave("images/sub/densities_frei.png")
```

There are many NA values, meaning that statistical power might be low when analysed at this granularity. Still, a striking result is that Strassenbau again shows a big structural break around year 6. The other sectors look like they could be collusive, but there is not a similarly clear breaking point.

```
# Look for break points within the Einladung procedure
# Number of bids
avg_var_year(df_einladung,"no_bids") +labs(y = "", x = "Year", title = "Mean number of bids per year",
       subtitle = "Among tenders of procedure Einladung") +coord_cartesian(ylim = c(0, 7))
```



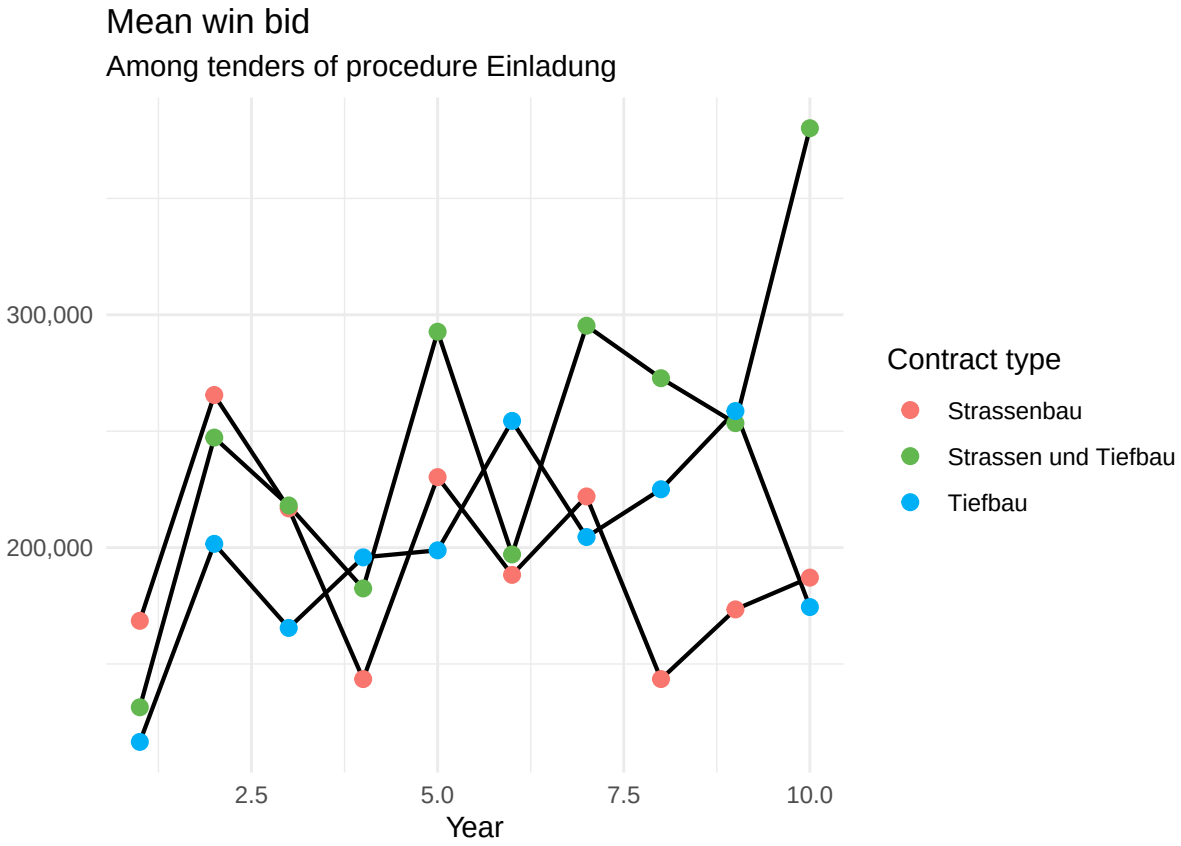

```
ggsave("images/sub/no_bids_ein_means.png", width = 24, height = 14, units = "cm")

box_plot(df_einladung, "no_bids", 1, "#f8766d", "Strassenbau and Einladung procedure", 0, 8, " ")
```



The mean number of bids look stable, but note that the average is lower than for the full sample. We know that fewer bidders is associated with a higher likelihood of a cartel. The values are also fairly stable over time with no indication of things changing at a particular point in time.

What about the winning bids?

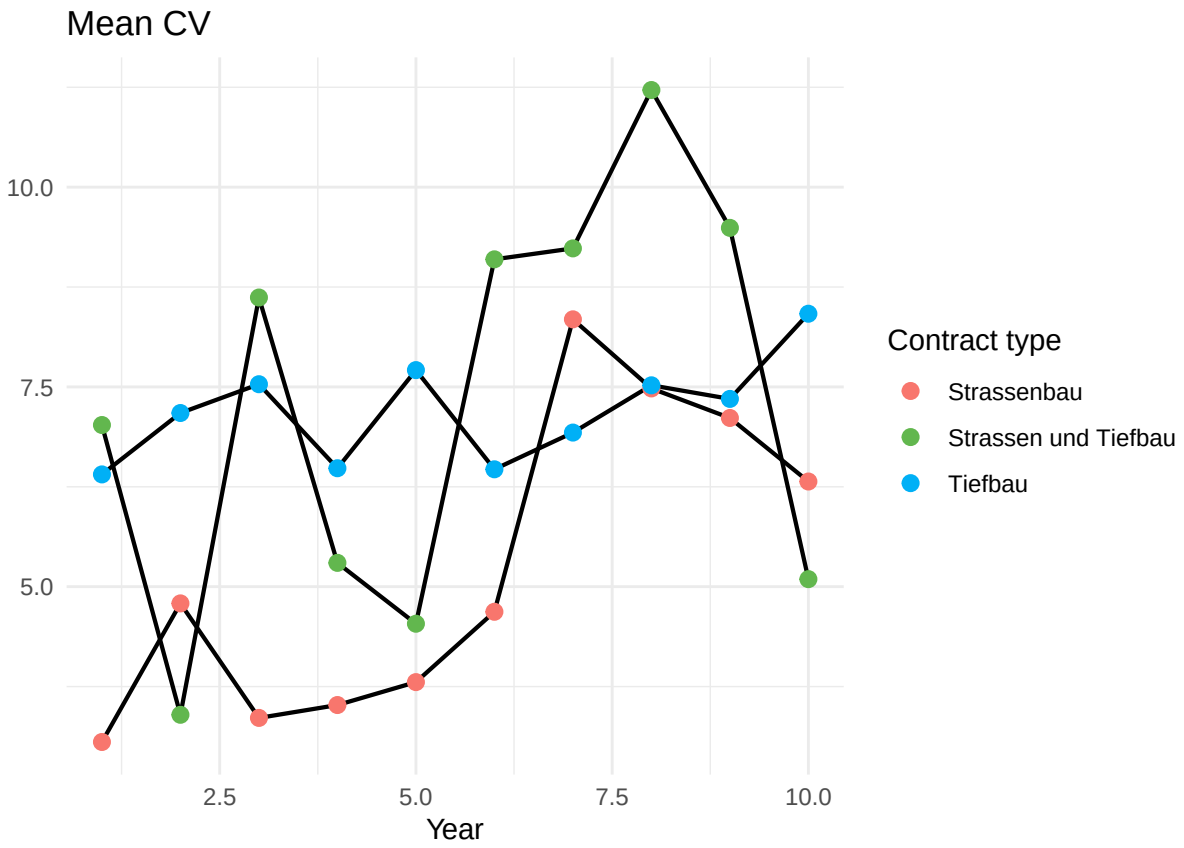
```
df_einladung %>% group_by(year, contract_type) %>%
summarise(mean_win_bid = mean(win_bid)) %>%
ggplot(aes(x = year, y = mean_win_bid)) +
  geom_line(size = 0.75, aes(fill = as.factor(contract_type))) +
  geom_point(size = 2.5, aes(color=factor(contract_type))) +
  theme_minimal() +
  labs(y = "CHF", x = "Year", title = "Mean win bid",
       subtitle = "Among tenders of procedure Einladung")  +
  scale_color_manual(name = "Contract type", values=c("#f8766d", "#62B74E", "#00b0f6"),
                     labels = c("Strassenbau", "Strassen und Tiefbau", "Tiefbau")) +
  scale_y_continuous(name="", labels = scales::comma)
```




```
ggsave("images/sub/no_bids_ein_wins.png", width = 24, height = 14, units = "cm")
```

Now, let's look at CV

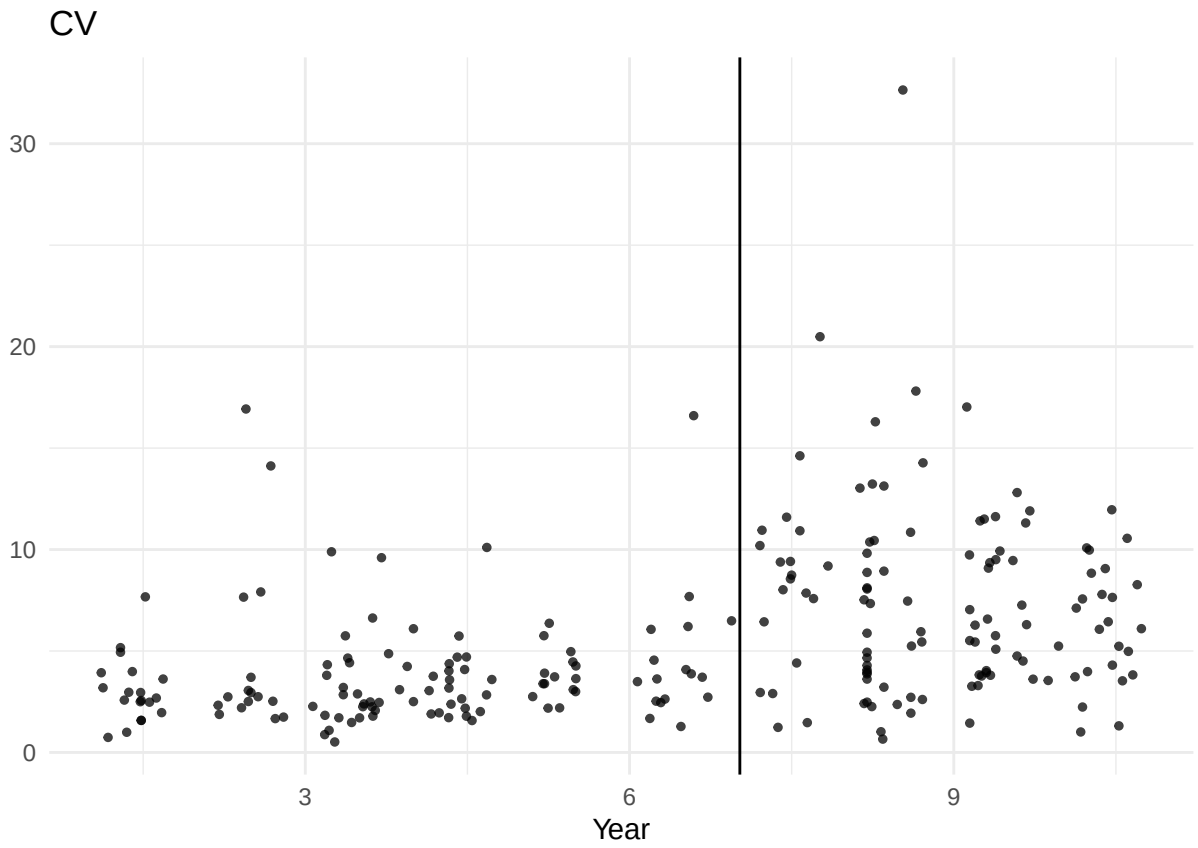
```
# Now, check CV:
#CV
df_einladung %>% group_by(year, contract_type) %>%
summarise(mean_CV = mean(CV)) %>%
ggplot(aes(x = year, y = mean_CV)) +
  geom_line(size = 0.75, aes(fill = as.factor(contract_type))) +
  geom_point(size = 2.5, aes(color=factor(contract_type))) +
  theme_minimal() +
  labs(y = "", x = "Year", title = "Mean CV")  +
  scale_color_manual(name = "Contract type", values=c("#f8766d", "#62B74E", "#00b0f6"),
                     labels = c("Strassenbau", "Strassen und Tiefbau", "Tiefbau"))
```




```
ggsave("images/sub/CV_sub.png", width = 24, height = 14, units = "cm")
```

There is a clear shift in the variable around year 6. Let's look at the raw data:

```
# Mann Whitney test scatter plot of break in CV for Strassenbau
bp <- 2200 # when using the "date" variable
bp2 <- (bp-1)/(365.25) + 1 # when using the date2 variable which shows values in years

df_einladung %>% filter(contract_type == 1) %>%
ggplot(aes(x = date2, y = CV)) +
  geom_point(size = 1, alpha = 0.75) +
  theme_minimal() +
  geom_vline(xintercept = bp2) +
  labs(y = "", x = "Year", title = "CV")
```




```
ggsave("images/sub/CV_point_str.png", width = 24, height = 14, units = "cm")


df_einladung_split <-df_einladung %>% mutate(pos = ifelse(date < bp, 1, 0))
wilcox.test(CV ~ pos, data=df_einladung_split) # Highly significant
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  CV by pos
## W = 70248, p-value = 2.112e-05
## alternative hypothesis: true location shift is not equal to 0
```

```
kstest_contract(df_einladung_split, 1)
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  x
## D = 0.11362, p-value = 0.09027
## alternative hypothesis: two-sided
```

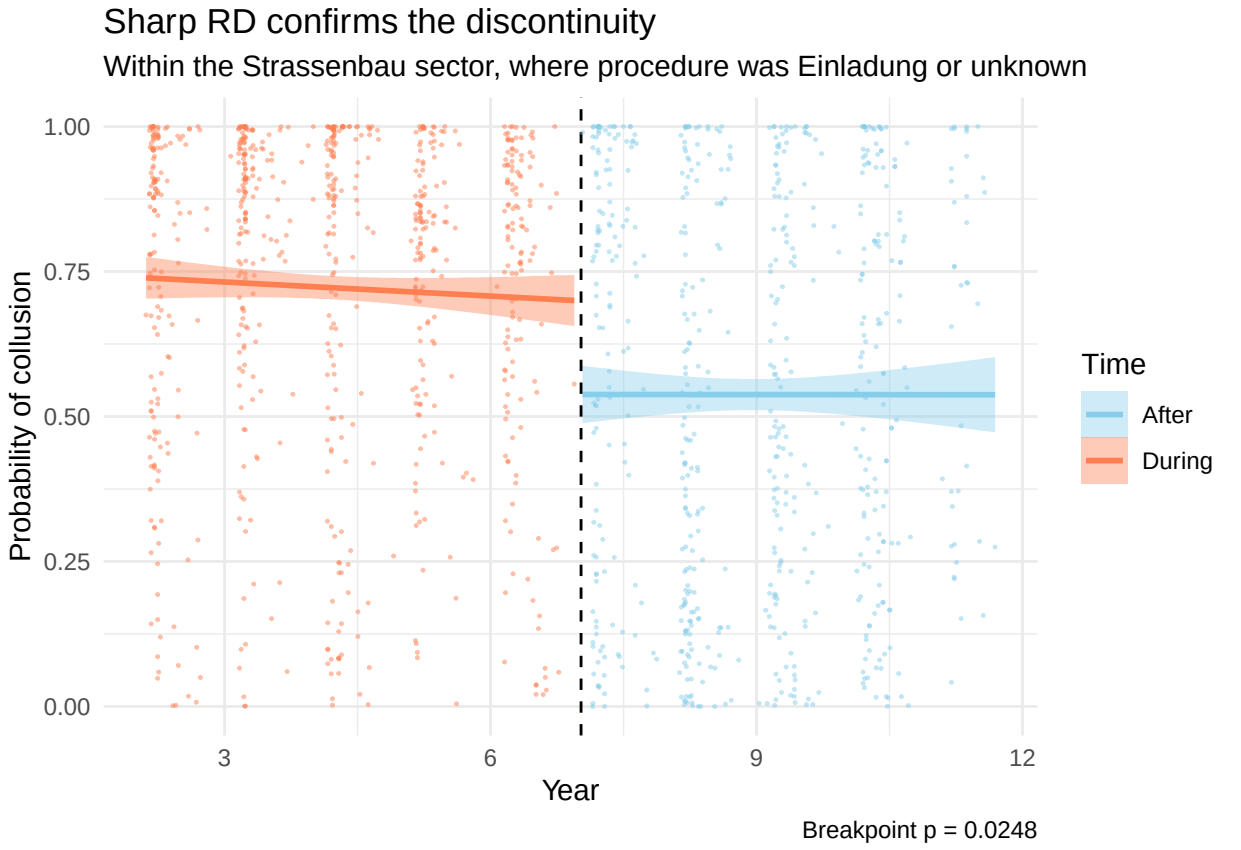
There is a visible change and the Mann Whitney U test statistic is significant meaning that there is a difference in medians before and after the breka point. Of course this might be due to a trend, but it looks like this is not the case.

To complete the analysis, we provide a regression discontinuity graph.

```r
# Add a rdd
df_rd <- df_agg %>% filter(contract_type == 1, procedure == "Einladung" | is.na(procedure)) %>%
  mutate(Treatment = as.factor(ifelse(date2 < bp2, 1, 0))) %>% filter(abs(date2 - bp2) < 5) # Apply fil

ols_fit <- lm(probability_collusion ~ date2 + Treatment*date2 + Treatment, data = df_rd)

summary(ols_fit)
```

```
##
## Call:
## lm(formula = probability_collusion ~ date2 + Treatment * date2 +
##     Treatment, data = df_rd)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.73515 -0.25611  0.08365  0.24972  0.46242
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        5.385e-01  9.015e-02   5.973 2.95e-09 ***
## date2             -8.194e-05  9.965e-03  -0.008   0.9934
## Treatment1         2.177e-01  9.693e-02   2.246   0.0248 *
## date2:Treatment1 -8.010e-03  1.280e-02  -0.626   0.5316
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3123 on 1394 degrees of freedom
## Multiple R-squared:  0.08051,    Adjusted R-squared:  0.07853
## F-statistic: 40.68 on 3 and 1394 DF,  p-value: < 2.2e-16
```

```r
p_rd <- summary(ols_fit)$coefficients["Treatment1",4]

df_rd %>%
  ggplot(aes(x = date2, y = probability_collusion,
             color = Treatment, fill = Treatment)) +
  geom_smooth(method = lm) +
  geom_vline(xintercept = bp2, linetype = "dashed") +
  theme_minimal() +
  geom_point(size = 0.2, alpha = 0.5) +
  scale_color_manual(name = "Time", values=c("skyblue", "coral"), labels = c("After", "During")) +
  scale_fill_manual(name = "Time", values=c("skyblue", "coral"), labels = c("After", "During")) +
  labs(x = "Year", y = "Probability of collusion",
       title = "Sharp RD confirms the discontinuity", subtitle = "Within the Strassenbau sector, where p
       caption = glue::glue("Breakpoint p = ", round(p_rd, 4)))
```




```
ggsave("images/strbau_rd.png", width = 24, height = 14, units = "cm")
# Note that graph is probably biased by the long estimation window and linearity
```

There is a significant break. However, there is a substantial degree of freedome here as the estimation window is somewhat arbitrary.

In addition, we've tried different Chow tests but since the break point is somewhat fluid, it doesn't yield any results. If the cartel breaks down gradually, the test may not be able to detect it.