sustainability

MDPI

File S1

This appendix presents the computer code that implements the Monte Carlo method described in “Confidence in greenhouse gas emission estimation: a case study of formaldehyde manufacturing”.

```
# YDK3.R
# Estimation of the mean and standard deviation of total direct emissions in a formaldehyde plant
# Monte Carlo computation of mu-hat, sigma-hat, and CI of year emission

setwd("C:/Users/ITA/Dropbox/DeepESG/Copenor")
rm(list=ls()) #Removes all items from the environment!
set.seed(123)

z <- 1
n <- 100000
CO2_mass <- c(rep(0,n))
kk <- (7216*273.15*44/10^(6)/134.4/0.022414) # formula constant

for (i in 1:n) {
  pp <- rnorm(1,mean=111.97  ,sd=z*1.57)
  ut <- rnorm(1,mean=0.00254,sd=z*0.00006)
  hh <- rnorm(1,mean=8278  ,sd=z*83) # anual
#  hh <- rnorm(1,mean=696  ,sd=z*83) # month of jan 2021
  vv <- (0.0144)-1*(0.0029)*(ut-0.00254)/0.00006  # vv correlated (rho = -1) with ut
# vv <- rnorm(1,mean=0.0144, sd=0.0029) # vv considered independent

  CO2_mass[i] <- kk*pp*ut*hh*vv
}
hist(CO2_mass)
mean(CO2_mass)
sd(CO2_mass)
# annual emission
print(c('95% CI for Annual emission CO2 =', mean(CO2_mass),' +/- ', 1.96*sd(CO2_mass)))
# the end
```

File S2

In this appendix, we discuss the effect of the normality assumption used in the article entitled "Confidence in greenhouse gas emission estimation: a case study of formaldehyde manufacturing".

We analyze the validity of the normality assumption for the variables *h*, *P*, *V*, and (*1*/*T*).

The total number of production hours, h, is subject to the variability introduced by the practice of recording some hours actually worked in a given year in the subsequent year (and vice versa). Other deviations in *h* arise from production interruptions and ramp-up periods. The combination of all these deviations was assumed to follow a normal distribution by means of the Central Limit Theorem. One must realize that no historical data on h would serve to determine its distribution since the annual production might vary substantially because of managerial decisions.

The variable (*1*/*T*) was assumed to be normally distributed on the basis of expert opinions. We consider that, in future research, the behavior of (*1*/*T*) could be investigated using data to be collected at appropriate points and at specific time intervals when other information are gathered simultaneously.

For the *V* and *P* values, we analyzed the classical QQ-plots, also called probability plots, [27] and performed the Shapiro–Wilk test of normality [28].

The graph in Figure S1 shows the QQ-plot of *V*. Visually, one may think that the cumulative distribution of V in the sample corresponds to the distribution of a normal variable with the same mean and variance. However, when we run the Shapiro–Wilk test, we find p-value = 0.03. Since this value is below 0.05, we refuse the normality hypothesis.

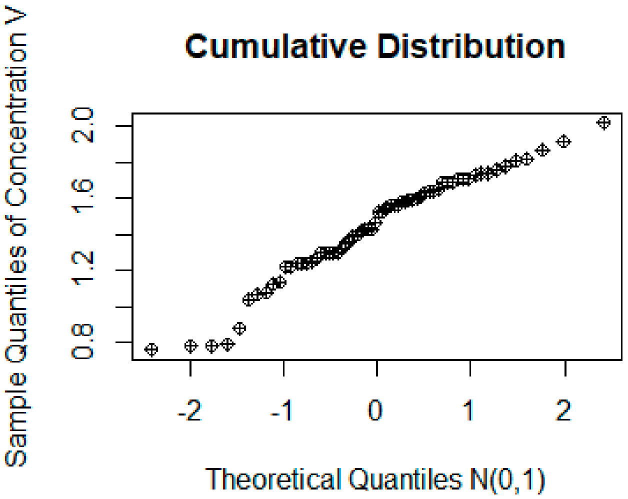

Figure S1: QQ-plot to test normality of *V*.

For the variable *P*, we also produced its QQ-plot, as shown in Figure S2. In this case, the cumulative distribution of P in the sample clearly does not correspond to a straight line because of the important deviation in both extremes. The Shapiro–Wilk test resulted in p-value = 5e-9. Since this value is below 0.05, we reject the normality hypothesis for p.

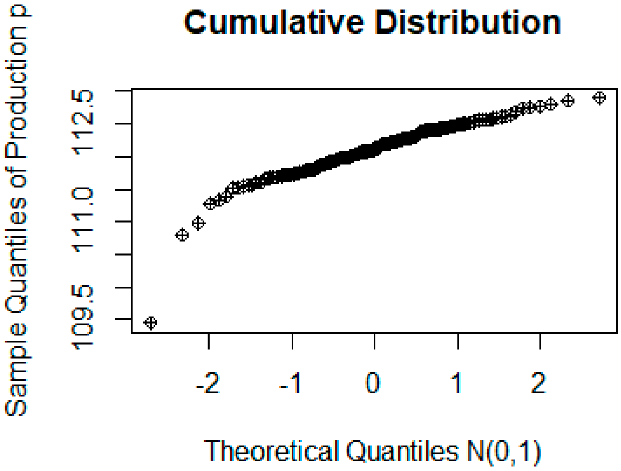

Figure S2: QQ-plot to test the normality of *P*.

We conclude that it would be more appropriate not to assume normal distributions to generate instances of variables V and P in the Monte Carlo simulations.

A non-parametric bootstrap method was implemented in an extension of the YDK.R algorithm that we call YDKboot.R. It is interesting to compare the results of the two models in the following table.

Table S1. Comparing the model using normal assumptions with the model using bootstrapping.

Item	YDK3.R Uses normality assumption	YDKboot.R Uses bootstrapping to simulate P and V
Considers correlation	Yes	Yes
Estimated annual emission	970	992
StdDev of annual emission	175	196
C.I. of annual emission	970 ± 342	992 ± 384

Using bootstrapping, the resulting estimate of total emissions during the year increased from 970 to 992. There are two approximations to explain such a difference. The first is simply the approximation introduced when a set of data points are approximated by a normal distribution. The second is related to the generation of correlated values. With the bootstrap procedure that we used, the correlation of extreme points assumed more importance than in the method using normal approximations.

When we do not use the normality assumption, we do not smooth the distribution of the input variables. As a consequence, the resulting histogram of the Monte Carlo estimation for the annual emission does not resemble a bell curve (Figure S3). As we can see below, there is a concentration of points in the lower extreme of the histogram, and the central region of the diagram is almost flat.

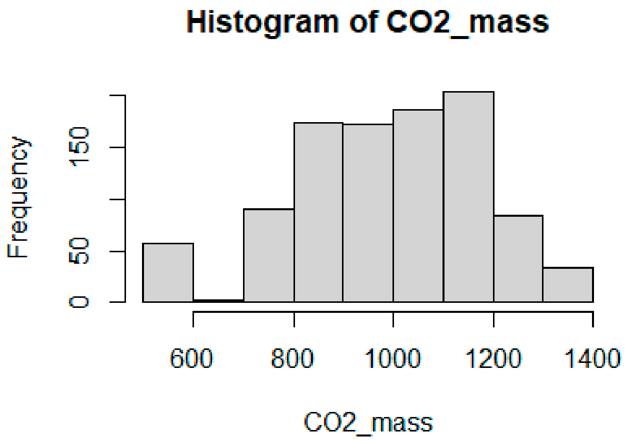

Figure S3: Histogram of simulated CO_2 mass emitted at the Copenor oxide unit in 2021.

References

1. Devore, J.L. *Probability and Statistics for Engineering and the Sciences*, 8th ed.; Cengage Learning: Belmont, CA, USA, 2012.
2. Wiki R Contributors. Teste Shapiro-Wilk (Ryan-Joiner). 2021. Available online: https://www.ufrgs.br/wiki-r/index.php?title=Teste_Shapiro-Wilk_(Ryan-Joiner)&oldid=3247 (accessed on 20 February 2023).