Mainstreaming GTAPBIO Database Build Process: Preliminary Version

I. Introduction

The GTAPBIO database and model have been used in several important studies regarding the economic outcomes as well as land use change impacts and associated carbon emissions from US biofuels policies and from alternative sustainable aviation fuel pathways (Chen et al 2018, Prussi et al 2021, Taheripour et al 2010, Zhao et al 2021). To improve the transparency and replicability of the GTAPBIO database, initial efforts have been made to mainstream its build process as part of the GTAP satellite databases. Combining existing database pipelines and scripts into a single workflow can help improve the timeliness of database release, provide better documentation and facilitate database improvements. This documentation outlines a preliminary workflow for creating an initial input to the GTAPBIO build process –an updated GTAP database which contains additional crops, vegetable oils, food and feed, electricity and power sectors. Note that this is an ongoing effort and revisions in the workflow and documentation are likely.

This build process is divided into three stages with default inputs and outputs provided in each separate folder of the replication zip archive[1]. Note that each stage should be implemented sequentially. In the first stage (folder '01_calc_target'), production and bilateral trade targets are constructed for each country given the new crops, vegetable oils and food and feed sectors using data from FAOSTAT (2026) as well as data from other external sources. These targets are then used to adjust the value of production and trade flow data in the initial version of the GTAP database with additional sector splits. These updated production and trade flow data are then used as inputs in the next stages. The second stage (folder '02_SAM_adjust') adjusts an aggregated GTAP social accounting matrix (SAM) - created from the initial version of the GTAP database with additional sector splits - using updated production and trade flow data from the first stage. Once the SAM accounts are adjusted, these are then used in the last stage (folder '03_database'). The last stage contains codes and data used to create the updated GTAP database

[1] The full replication zip archive and Github links are available upon request. Note that Github has file size limits so only certain files are hosted in the website. However, a separate link for a full replication zip archive which contains all the files and input data are provided in the Github link.

with new sectors given inputs from the previous stages. Additional files are also created from the updated database for viewing key flows and economic accounts as well as tax rates.

To run the whole build process, users are required to install the following software tools in their computer:

1) R statistical software (https://www.r-project.org/) (see *.r files) with the following packages installed:
    a. 'tidyverse' for data processing
    b. 'HARr' for writing header array files used by GEMPACK (*.har files)
    c. 'doParallel' and 'foreach' packages for parallel processing to reduce computation time
2) R Studio software for running the R codes (https://posit.co/download/rstudio-desktop/)
3) GAMS software (https://www.gams.com/) with CONOPT license for running cross entropy algorithm when balancing SAM accounts (see *.gms and *.gdx files)
4) GEMPACK program and license (https://www.copsmodels.com/gempack.htm) for running the rest of the data build codes (see *.tab, *.cmf and *.har files)

This preliminary build process has been tested under Windows operating system and using R studio software to run the main R files for each stage of the database build. In each stage, there is a "run_all.r" file which is designed to run all scripts within each stage sequentially via the R environment. Users are then required to manually copy outputs ('out' folder) as inputs files ('in' folder) in other stages. These include "targets.har" which is a header array file that contains the updated production and trade flows produced from the first stage – used as an input in the second and the last stages – as well as "balreg.har" which is another header array file that contains the adjusted aggregated GTAP social accounting matrix and is used as an input in the last stage.

II. Detailed description of each stage in the build process

1. Updating Production and Bilateral Trade Flows. All the files used in the first stage are in the '01_calc_target' folder (Figure 1). Inputs used in this stage include value of agricultural production, aggregate and bilateral export and import flows from FAOSTAT (2026) (vprod.zip,

trad.zip, tradmat.zip), additional production data for vegetable oilseeds sector from other external sources (vol_production_data.csv), as well as an initial GTAP database that has additional crops, vegetable oils, food and feed, electricity and power sectors (basedata.har, baseview.har, baserate.har, default.prm, sets.har). In this version of the build, the initial GTAP database is based on version 11a for year 2017 (Aguiar et al 2023) with energy sectors from the GTAP Power database (Chepeliev 2023) as well as additional sectors such as corn, sorghum and other coarse grains (from coarse grains sector); soybean, palm oil, rapeseed and other oilseeds (from oilseed sector); vegetable oil soy, vegetable oil palm, vegetable oil rapeseed, other vegetable oil sector (from vegetable oil sector) and food and feed sectors (from food sector) (Tables 1 and 2). There are several set mapping files to trace countries across different country ISO Codes, FAO Area Codes and GTAP regions (country_sets.csv) as well as to track sectors across different CPC commodity codes into GTAP sectors (agg_trade_comm_sets.csv, detailed_trade_comm_sets.csv, value_comm_sets.csv). There are also additional files which define the order of region and commodity elements (gtap_sets.csv) as well as new sector sets (gtapbiosets.har).

The main R code needed to run all the processes in this stage is the “run_all.r” file. The sequence of the build process is as follows. First, the required packages are loaded in the R environment and temporary and output directories are created. Users then have the option to download FAOSTAT data within R or use a version of the data stored in the ‘in’ folder (the latter is the default option). After defining the year for data processing (i.e. 2017), country and commodity set files, additional production data for the vegetable oil sectors are read in the R environment. In the past versions of GTAPBIO, this additional data comes from World Oil database (Taheripour et al 2017). Once the input files are loaded in the R environment, the code then runs a series of data processes to attach country and commodity set mapping to the raw data, aggregate the raw data to GTAP regions and sectors and write the data in header array formats used by GEMPACK software. Once created, these header array files are then used to adjust the production and trade flow data in the initial version of the GTAP database with additional sector splits (hi.tab). To balance the adjusted bilateral trade flows, the RAS algorithm is used. After this process, the final output of the first stage is produced which is a header array file containing the updated data for the value of production, as well as bilateral import and export flows at market prices and world prices (see targets.har). Users need to copy and transfer this file in the input folders of the next stages of this workflow.

2. Adjusting aggregated GTAP social accounting matrix. The main purpose of the second stage is to adjust the aggregated GTAP social accounting matrix (SAM) accounts using the updated production and trade flows created from the first stage. The inputs in this stage include GDP data from GTAP database (baseview.har), aggregate and disaggregate GTAP SAM accounts (gtapsam.har) - based on the initial GTAP database with additional sector splits that was used in the first stage - as well as set files which contain GTAP region sets (reg_set.csv) as well as sector and endowment sets (set.txt). Note that the output of the first stage is also used as an input in the second stage (targets.har). This stage also contains executable files[2] to convert inputs and outputs across different file formats. These include 'txt2har.exe' which converts text files into header array format used by GEMPACK software as well as 'gdx2har.exe' and 'har2gdx.exe' to convert gdx files produced by GAMS software into header array used by GEMPACK and vice-versa.

Similar to the first stage, users need to run the 'run_all.r' file in the R environment which implements codes in this stage sequentially. There are five steps in this stage. The first step (run_isresh.r) runs the adjustment program (isresh.tab) which reads in the disaggregated SAM accounts, creates initial aggregated SAM accounts and adjusts the aggregated SAM based on the updated production and trade flow data from the first stage. After adjusting aggregated SAM accounts, these outputs are converted from har format into gdx format. This gdx file is read in GAMS software in the next step (run_cesam2.r) which further adjusts the aggregated SAM accounts for each country using cross entropy algorithm. Note that parallel processing is implemented in the R code in order to implement adjustments by country simultaneously[3]. After these adjustments, the third step (run_regproc.r) creates a reporting file (balstatus.gdx) while the fourth step (run_resconcat.r) runs the program that aggregates individual SAM by country into a single gdx file (balreg.gdx) which is then converted into har file format (balreg.har). This is the main output of the second stage and users need to copy and transfer this file as an input to the final stage. The last step (run_postbal.r) creates a reporting file which compares the initial and adjusted aggregated SAM accounts (sambalreg.har).

[2] These executable files are provided in the zip archive.
[3] The default version uses 8 cores but users can change this number via the registerDoParallel() command in R

3. Creating the updated GTAP database. The last stage of the build process uses the initial GTAP database with sector splits used in the first stage (basedata.har), updated data for the value of production and trade flows also from the first stage (targets.har), as well as adjusted aggregated SAM accounts from the second stage (balreg.har) to create the updated GTAP database. In general, there are 4 steps in this stage and these are implemented together using the "run_all.r" file within the R environment. The first step (oshares.tab) calculates key domestic, intermediate and final uses, import and margin shares as well as export and import tax rates from the initial GTAP database. This information is used for checking and used as inputs in the next steps. The second step (cmbtrd.tab) combines the trade flow information from the initial GTAP database, for sectors which have not been adjusted, with the updated trade flow data generated in the first stage for selected sectors. The third step (convert.tab) makes the updated GTAP database (basedata.har) – the final output of this preliminary build process. It converts the adjusted aggregated SAM accounts into GTAP database format. It also includes several manual adjustments for special cases such as negative flows and low use for commodities. After these adjustments, the updated GTAP database is constructed using information from the adjusted aggregated SAM accounts as well as the initial GTAP database. The final step (gtapview.tab) creates additional files for viewing key flows and economic accounts as well as tax rates (baseview.har, baserate.har) from the updated database.

III. Summary and Conclusion

To summarize, this documentation provides an overview of the preliminary workflow for creating an initial input to the GTAPBIO build process – an updated GTAP database which contains additional crops, vegetable oils, food and feed, electricity and power sectors. This preliminary workflow has been constructed using existing database pipelines and codes and combines them in a single build process to improve transparency and replicability. In the future, revisions in the workflow and documentation are needed as the database build is further improved and optimized. More importantly, this workflow covers only a small portion of the full GTAPBIO database build and future work should also focus on incorporating the rest of the database build.

References

Aguiar, A., Chepeliev, M., Corong, E., & van der Mensbrugghe, D. (2023). The Global Trade Analysis Project (GTAP) Data Base: Version 11. *Journal of Global Economic Analysis*, *7*(2). https://doi.org/10.21642/JGEA.070201AF

Chen, R., Qin, Z., Han, J., Wang, M., Taheripour, F., Tyner, W., O'Connor, D., & Duffield, J. (2018). Life cycle energy and greenhouse gas emission effects of biodiesel in the United States with induced land use change impacts. *Bioresource Technology*, *251*, 249–258. https://doi.org/https://doi.org/10.1016/j.biortech.2017.12.031

Chepeliev, M. (2023). GTAP-Power Data Base: Version 11. *Journal of Global Economic Analysis*, *8*(2). https://doi.org/10.21642/JGEA.080203AF

FAOSTAT Statistical Database (2026). Rome, Italy. Accessed March 2026. https://www.fao.org/faostat/en/.

Prussi, M., Lee, U., Wang, M., Malina, R., Valin, H., Taheripour, F., Velarde, C., Staples, M. D., Lonza, L., & Hileman, J. I. (2021). CORSIA: The first internationally adopted approach to calculate life-cycle GHG emissions for aviation fuels. *Renewable and Sustainable Energy Reviews*, *150*, 111398. https://doi.org/https://doi.org/10.1016/j.rser.2021.111398

Taheripour, F., Hertel, T. W., Tyner, W. E., Beckman, J. F., & Birur, D. K. (2010). Biofuels and their by-products: Global economic and environmental implications. *Biomass and Bioenergy*, *34*(3), 278–289. https://doi.org/https://doi.org/10.1016/j.biombioe.2009.10.017

Taheripour, F., Pena-Levano, L., & Tyner, W. (2017). *Introducing first and second generation biofuels into GTAP 9 Data Base* (GTAP Research Memorandum No. 29). Purdue University, West Lafayette, IN: Global Trade Analysis Project (GTAP)

Zhao, X., Taheripour, F., Malina, R., Staples, M. D., & Tyner, W. E. (2021). Estimating induced land use change emissions for sustainable aviation biofuel pathways. *Science of The Total Environment*, *779*, 146238. https://doi.org/https://doi.org/10.1016/j.scitotenv.2021.146238

Table 1. GTAP region codes (n = 160)

1	aus	41	ven	81	esp	121	xnf
2	nzl	42	xsm	82	swe	122	ben
3	xoc	43	cri	83	gbr	123	bfa
4	chn	44	gtm	84	che	124	cmr
5	hkg	45	hnd	85	nor	125	civ
6	jpn	46	nic	86	xef	126	gha
7	kor	47	pan	87	alb	127	gin
8	mng	48	slv	88	srb	128	mli
9	twn	49	xca	89	blr	129	ner
10	xea	50	dom	90	rus	130	nga
11	brn	51	hti	91	ukr	131	sen
12	khm	52	jam	92	xee	132	tgo
13	idn	53	pri	93	xer	133	xwf
14	lao	54	tto	94	kaz	134	caf
15	mys	55	xcb	95	kgz	135	tcd
16	phl	56	aut	96	tjk	136	cog
17	sgp	57	bel	97	uzb	137	cod
18	tha	58	bgr	98	xsu	138	gnq
19	vnm	59	hrv	99	arm	139	gab
20	xse	60	cyp	100	aze	140	xac
21	afg	61	cze	101	geo	141	com
22	bgd	62	dnk	102	bhr	142	eth
23	ind	63	est	103	irn	143	ken
24	npl	64	fin	104	irq	144	mdg
25	pak	65	fra	105	isr	145	mwi
26	lka	66	deu	106	jor	146	mus
27	xsa	67	grc	107	kwt	147	moz
28	can	68	hun	108	lbn	148	rwa
29	usa	69	irl	109	omn	149	sdn
30	mex	70	ita	110	pse	150	tza
31	xna	71	lva	111	qat	151	uga
32	arg	72	ltu	112	sau	152	zmb
33	bol	73	lux	113	syr	153	zwe
34	bra	74	mlt	114	tur	154	xec
35	chl	75	nld	115	are	155	bwa
36	col	76	pol	116	xws	156	swz
37	ecu	77	prt	117	dza	157	nam
38	pry	78	rou	118	egy	158	zaf
39	per	79	svk	119	mar	159	xsc
40	ury	80	svn	120	tun	160	xtw

Table 2. GTAP sector codes (n = 86)

1	pdr	36	tex	71	trd
2	wht	37	wap	72	afs
3	Corn	38	lea	73	otp
4	Sorghum	39	lum	74	wtp
5	Oth_CrGr	40	ppp	75	atp
6	v_f	41	p_c1	76	whs
7	Soybeans	42	Blender	77	cmn
8	Palm_f	43	chm	78	ofi
9	Rapeseed	44	bph	79	ins
10	Oth_Oilseeds	45	rpp	80	rsa
11	c_b	46	nmm	81	obs
12	pfb	47	i_s	82	ros
13	ocr	48	nfm	83	osg
14	ctl	49	fmp	84	edu
15	oap	50	ele	85	hht
16	rmk	51	eeq	86	dwe
17	wol	52	ome		
18	frs	53	mvh		
19	fsh	54	otn		
20	coa	55	omf		
21	oil	56	TnD		
22	gas	57	NuclearBL		
23	oxt	58	CoalBL		
24	cmt	59	GasBL		
25	omt	60	WindBL		
26	vol_soy	61	HydroBL		
27	vol_Plam	62	OilBL		
28	vol_Rape	63	OtherBL		
29	vol_othn	64	GasP		
30	mil	65	HydroP		
31	pcr	66	OilP		
32	sgr	67	SolarP		
33	Food	68	gdt		
34	Feed	69	wtr		
35	b_t	70	cns		

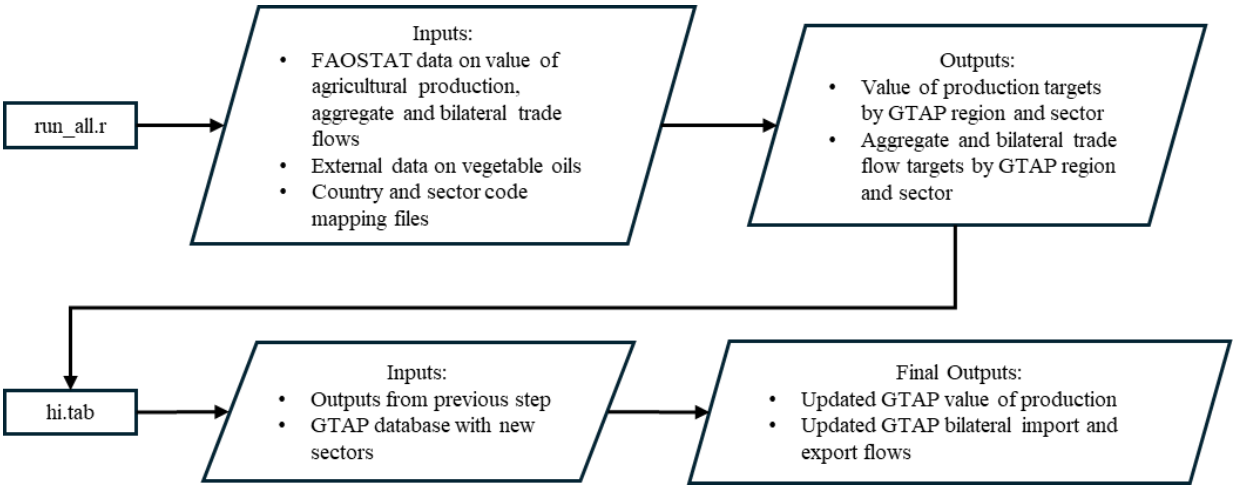

Figure 1. Overview of the First Stage: Updating Production and Bilateral Trade Flows.

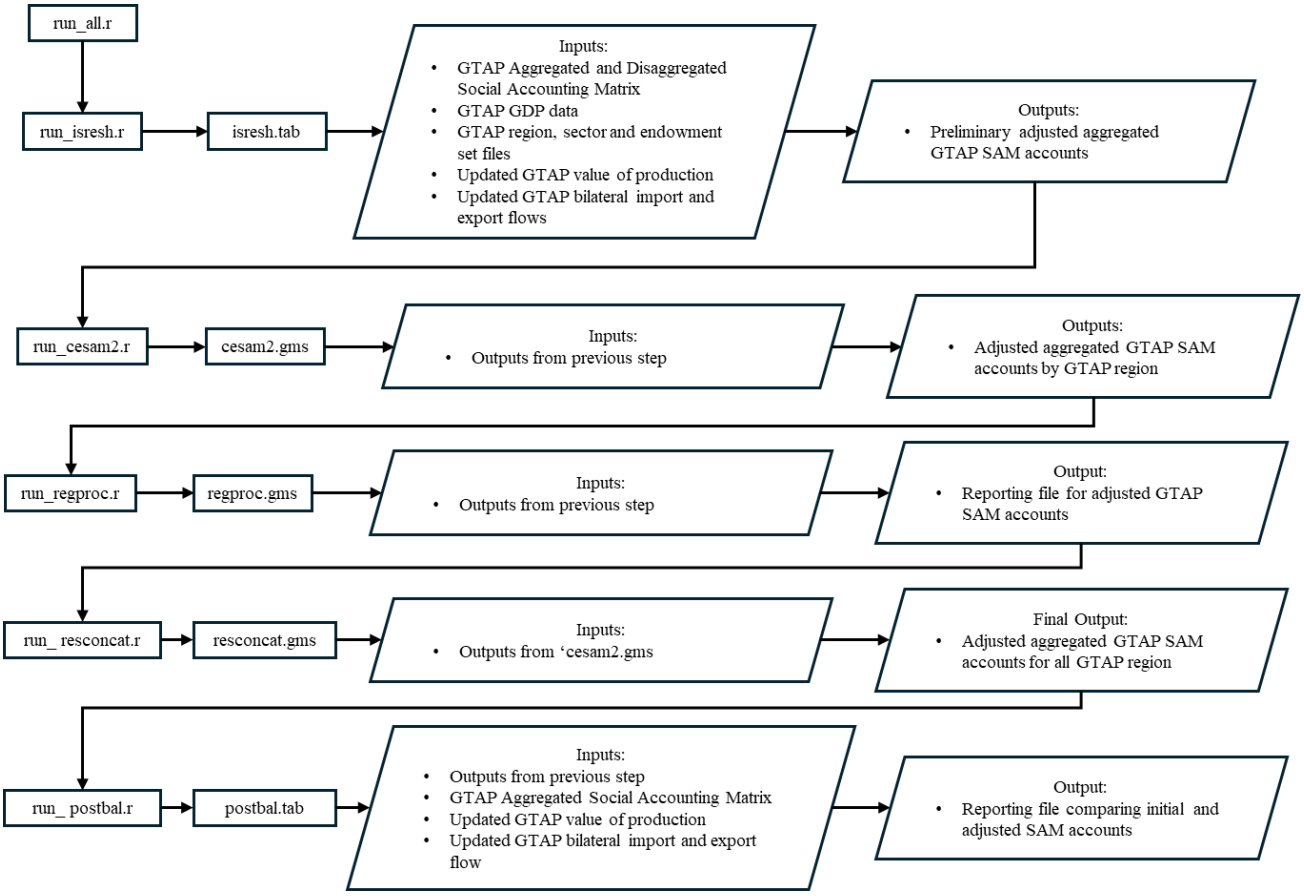

Figure 2. Overview of the Second Stage: Adjusting aggregated GTAP social accounting matrix

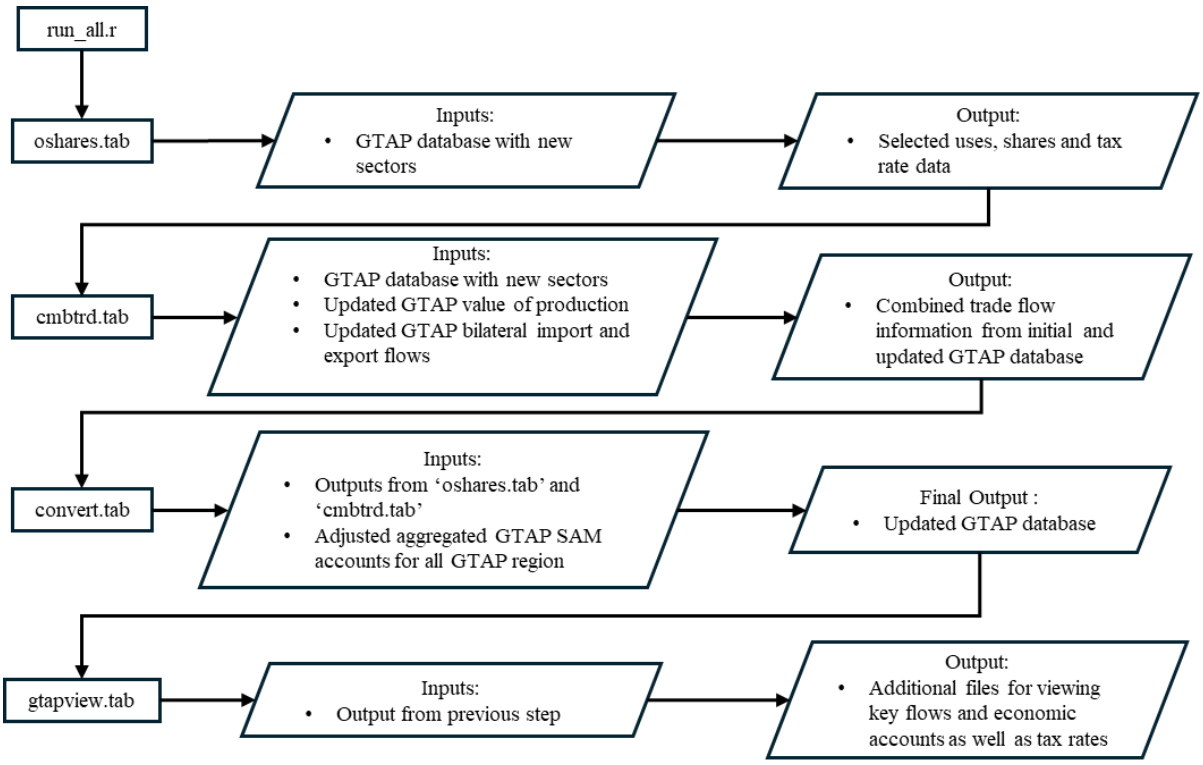

Figure 3. Overview of the Last Stage: Creating the updated GTAP database