# Virtual Beach 3.0.6 – Data Preparation for GBM model

#### In this module you will learn how to:

- A. Import and clean-up model-building data for your beach
- B. Process directional data (wind, currents, waves)
- C. Combine two or more predictive variables
- D. Transform the dependent bacteria variable

#### A. Import and clean-up model-building data for your beach

A.1. Open **Microsoft Excel** to preview the data you will be importing into **Virtual Beach 3**. Open the file "VB\_Training\_Data\_GBM.xls".

Be sure to save your data as "\*.xls" files. A plugin for **Virtual Beach** 3 is available for importing "\*.xlsx" files, but there are still bugs to be worked out.

Column **B** is always the *response* variable, "ECOLI" in this example. All data to the right are potential *explanatory* variable. See the **KEY** tab of the **Excel** file for descriptions of variables used in this module. Close the **Excel** file before returning to **Virtual Beach 3**. Data cannot be imported from an open Excel file.

G	<u>יי</u> ירי ד			VB_Tra	aining_Data_GBM.xls	[Compatibility Mo	de] - Excel		困 -		K
Fi	le Home Ins	ert F	Page Aayout	Formulas Da	ta Review Vi	ew ACROBAT	♀ Tell me wha	t you want to do	Sign ir	A Share	е
B1	• :	×	fx E	COLI							*
	А	В	С	D	E	F	G	Н	I.	J	
1	DATETIME	ECOLI	QTRSEASON	PRE_JUNE21	JUNE21_JULY15	JULY16_AUG10	POST_AUG10	CLOUDCOV_Qual	DOY	RRAIN6	
2	5/21/2009 12:05	2	1	. 1	0	0	0	Sunny	141		
3	5/28/2009 12:20	5	1	. 1	0	0	0	Cloudy	148	6	
4	6/4/2009 11:55	1	1	. 1	0	0	0	Sunny	155		
5	6/11/2009 12:35	345	1	. 1	0	0	0	Mostly Sunny	162		
6	6/12/2009 14:15	18	1	. 1	0	0	0	Cloudy	163		
7	6/15/2009 11:25	29	1	. 1	0	0	0	Sunny	166	j -	
8	6/16/2009 10:30	8	1	. 1	0	0	0	Cloudy	167	1	
9	6/17/2009 14:05	120	1	. 1	0	0	0	Sunny	168	•	
10	6/18/2009 14:05	17	1	. 1	0	0	0	Sunny	169	)	
11	6/22/2009 10:40	4	2	0	1	0	0	Sunny	173		
12	6/23/2009 11:45	76	2	0	1	0	0	Partly Sunny	174		
13	6/24/2000 55	15	2	0	1	0	0	Partly Sunny	175		-
4	▶ KEY	BM-M	odel-Buildin	g data_2009-1	GBM-Valida	tion data_2016	+				]

This file can be used as a template for formatting beach-specific data.

A.2. Return to Virtual Beach 3 project file created in the "Beach Orientation" module.
1. Click the Global Datasheet tab. 2. Click the Import Data icon and select the Excel file "VB\_Training\_Data\_GBM.xls". 3. Click Open.

Re Location Global Datasheet	n Open		×
0 0 2 2 3 0	This PC > OS (C:) > Virtual Beach 3	v ©	Search Virtual Beach 3
nport Validate Compute Manipulate Transform Go To Data Cata A.O Model	Organize - New folder		BE + 🔟 🕜
dd Vork with Data	Downloads ^ Name	Date modified	Type Size
2	Music Britures Videos Solutions Color VB_Training_Data_GBM.xls	-8/15/2017 6:13 PM	Microsoft Excel 97 579 KB
	File name: VB_Training_Data_GBM.xls	~	Sprevert Files (*.xls;*.xls;c*.c ~ Open Cancel

A.3. In this example there is more than one worksheet in the Excel file, so you must choose which one to import. **1.** Select the worksheet **GBM-Model-Building data\_2009-15. 2.** Click **OK**.

🖳 Virtual	Beach 3					
File	Location	Global	Datasheet			
0	0	34	24	3	0	Select Excel Workbook — 🗆 🗙
Import	Validate	Compute	Manipulate	Transform	Go To Model	Worksheets in VB_Training_Data_GBM.xls
Add	Validate	40	Work wit	h Data	Model	GBM-Model-Building data_2009-15 GBM-Validation data_2016
-				_	_	
						1
						2
						Ok Cancel

A.4. The data table will open in **Virtual Beach 3**.



**F** Virtual Beach 3 automatically highlights the second column of the datasheet as the response variable, "ECOLI", in this example. The "Response Variable" is indicated in the left-hand panel, along with "Column Count", "Row Count" and other descriptions of the data.

	Location Glob	al Datasheet									
ort a d	Validate Validate Validate	C ( te Maripulate Tra Work with Da	Som	Go Tu Model							
File		1-VB-Training-Data		DATETIME	ECOLI	QTRSEASON	PRE_JUNE21	JUNE21_JULY15	JULY16_AUG10	POST_AUG10	C
Colum	n Count	96		5/21/2009 12:05	2	1	1	0	0	0	Su
low C	w Count 286			5/28/2009 12:20	5	1	1	0	0	0	Clc
ate-	Time Index DATETIME		6/4/2009 11:55:	1	1	1	0	0	0	Su	
espu	inse vanable	ECOLI		6/11/2009 12:35	345	1	1	0	0	0	Mc
lisabl	ed Row Count	0		6/12/2009 2:15:	18	1	1	0	0	0	Ck
isabl	ed Column Count	0		6/15/2009 11:25	29	1	1	0	0	0	Su
iddei	n Column Count endent Variable Count	94		6/16/2009 10:30	8	1	1	0	0	0	Clc
				6/17/2009 2:05:	120	1	1	0	0	0	Su
				6/18/2009 2:05:	17	1	1	0	0	0	Su
				6/22/2009 10:40.,,	4	2	0	1	0	0	Su
			1	C /22 /2000 11-45	70	2	n.	1	n	n	P- *

A.5. Virtual Beach 3 will NOT build a model if any cells have null (missing), or nonnumeric (text) values. 1. Click the Validate Data icon to check your dataset. 2. In the pop-up window, click Scan.

🛃 Virtual Beach 3	Validation						-	
Import Add Validate	Data Validation Scan (Optional) Find:							
File	Action	ECOLI	QTRSEASON	PRE JUNE21	JUNE21 JULY15	JULY16 AUG10	POST AUG10	C
Column Count	(@) Replace With:	2	1	1	0	0	0	Su
Row Count	O Delete Row	5	1	1	0	0	0	Clc
Date-Time Index	O Delete Column	1	1	1	0	0	0	Su
Response Variable	Take Action Within	345	1	1	0	0	0	Mc
Disabled Row Count	Only This Cell 🗸	18	1	1	0	0	0	Clc
Disabled Column Count		29	1	1	0	0	0	Su
Hidden Column Count	Take Petron	8	1	1	0	0	0	Clc
independent valiable de		120	1	1	0	0	0	Su
		17	Ť	1	0	0	0	Su
	Identify Categorical Variables	4	2	0	1	0	0	Su
<		70	2	n	1	0	n	0-
Location Global Da	Return	-						_

A.6. In this example, the "CLOUDCOV\_qual" column is flagged because the values are text, or non-numeric. **1.** Click the radio button next to **Delete Column**. Under **Take Action Within** make sure **Only This Column** is selected. **2.** Click **Take Action**.

Filter Desello	vandauon						_	1
Location	Data Validation	-						
rt Validate Cor a Data i Validate	Scan (Optional) Find:							
le		QTRSEASON	PRE_JUNE21	JUNE21_JULY15	JULY16_AUG10	POST_AUG1	CLOUDCOV_Qual	
olumn Count	O Replace With:	1	1	0	0	0	Sunny	
w Count	O Delete Row	1	1	0	0	0	Cloudy	
te-Time Index	Delete Column	1	1	0	0	0		
sponse vanable	Take Action Within:	1	1	0	0	0	Mostly Sunny	1
abled Row Count	Only This Column 🗸	1	1	0	0	0	Cloudy	1
abled Column Count		1	1	0	0	0	Sunny	1
den Column Count enendent Variable Co	Take Action	1	1	0	0	0	Cloudy	1
spendent vanable of		1	1	0	0	0	Sunny	1
	-11	1	1	0	0	0	Sunny	1
	Identify Categorical Variables	2	0	1	0	0	Sunny	1
		4	n	1	n	n	Dauthe Comme	t
	time and the second second	-						-
	Febum Cancel							

A.7. Repeat step A.6 until you come to the "TRIB6" column. The variable is numeric, but some cells are empty. Do not remove the entire column. **1.** Click the radio button for **Delete Row**. **2.** Select **Entire Column** and click **Take Action**.

Virtual Beach 3	Validation						-	Ц (
Location Validate ta Data Idi Validate	Data Validation Scan (Optional) Find:							
File	Action:	RRAIN24	RRAIN48	RRAIN72	RRAIN120	RRAIN144	TRIB6	-
Column Count	Delete Row	0	0	0	60.27	63.52	813	9
Row Count	O Delete Column	0.88	0.88	0.88	12.76	62.9	571.7	6
Response Variable	C) boldte coldmit	0	0.88	0.88	88.0	12.76	586.5	5
	Take Action Within:	0	0	22.51	22.51	23.39	333.1	4
Disabled Row Count	Only This Row	0	0	0	22.51	22.51	193.3	2
Disabled Column Count	Only This Row	0.38	0.38	0.38	22.89	22.89	200.8	1
Hidden Column Count	Entire Sheet	0	0.38	0.38	0.38	22.89	206.2	2
	hr	1.63	12.52	12.52	12.52	12.9	169.6	1
		0	1.63	12.52	12.52	12.52	161.4	1
	Identify Categorical Variables	7 16	16	17.63	28.52	28.52		
							-	1.04

A.8. Repeat Step A.7 until a notice appears at the bottom of the pop-up window stating **No anomalous data values found**. Then click the **Return** button.

Location Validate Data Validate	Validation Data Validation Scan						
le	(Optional) Find:	ECOLI	QTRSEASON	PRE_JUNE21	JUNE21_JULY15	JULY16_AUG10	POST_AUG10
olumn Count	Action	345	1	1	0	0	0
ow Count	Benlace With:	. 18	1	1	0	0	0
ate-Time Index	(ii) Delete Row	29	1	1	0	0	0
esponse valiable	Datas Column	. 17	1	1	0	0	0
sabled Row Count	C. Delete Colorini	4	2	0	1	0	0
sabled Column Count	Take Action Within	76	2	0	1	0	0
dden Column Count	Entire Column	15	2	0	1	0	0
		3	2	0	1	0	0
	Tané na si	11	2	0	1	0	0
	^	. 5	2	0	1	0	0
	Identify Categorical Variables	. 7	2	0	1	0	0
		26	2	0	1	0	0
	No anomalous data values fou	nd.					2

### B. Process wind and current data

B.1. **1.** Click the **Compute A O** icon. **2.** In the pop-up window, under **Wind Data**, click the pull down arrow next to **Speed** and select WSPD. For **Direction**, select WDIR.

Virtual Beach 3	🖳 Wind/Current C	omponents – 🗆 🗙	-					
Location Global	Wind Data Specify wind data	columns:						
a Data AO d Valida	Speed Direction (deg)	WVPD24  WVPAR WVPERP WVPAR3 WVPERP3	OTRSEASON	PRE IIINE21	UNE21 10 Y15		POST AUG10	0.0
Column Count	Current Data	WVPAR6	1	1	0	0	0	14
ow Count 2	Conent Data	WVPAR12	1	1	0	0	0	14
ate-Time Index [	Specify current da	WVPERP12 WVPAR24	1	1	0	0	0	15
esponse Variable B	Speed	WVPERP24	1	1	0	0	0	16
sabled Row Count		WDIR	1	1	0	0	0	16
sabled Column Count	Direction (deg)	WPERP	1	1	0	0	0	16
dden Column Count (		WPAR3 WPERP3	1	1	0	0	0	16
dependent Variable Count	Wave Data	WPAR6	1	1	0	0	0	16
	Wave Data	WPERP6 WPAR12	1	1	0	0	0	16
	Specify wave data	WPERP12	2	0	1	0	0	17
	Height	WPERP24 ATEMP ATEMP6	2	n	1	0	n	17
ocation Global Datashe	Direction (deg)	ATEMP12 ATEMP24 ATEMPstdv6						

B.2. Repeat for **Current Data**, selecting CSPD and CDIR. Repeat for **Wave Data** selecting WVHT and WVDIR. The Beach Angle is automatically included. Click **OK**.

wind Data								
Specify wind data	columns:							
Speed	WSPD	~						
Direction (deg)	WDIR	~ RP3	CPAR6	CPERP6	CPAR12	CPERP12	CPAR24	CPERP24
		13	0.1261	-0.01659	0.1717	-0.03284	0.1625	-0.03638
Current Data		8	-0.1912	0.0718	-0.2873	0.09819	-0.437	0.1281
Specify current da	ta columns:	5	-0.102	0.03684	-0.1361	0.03951	-0.1784	0.04436
			-0.2051	0.04907	-0.1737	0.04087	-0.1289	0.02522
Speed	CSPD	× 13	-0.1331	0.04138	-0.146	0.04777	-0.1977	0.06138
Distant (dea)	CDIR	09	0.06425	-0.009418	0.09981	0.004245	0.1182	-0.01211
Direction (deg)	CDIR	1187	-0.1005	-0.01045	-0.08743	-0.004734	-0.01349	-0.01155
		09	-0.209	0.05336	-0.1797	0.04137	-0.1853	0.03237
Wave Data		57	-0.02085	0.001331	-0.04001	0.02431	-0.07802	0.02366
Specify wave data	a columns:	722	-0.0004242	-0.01033	-0.04925	-0.01564	-0.05114	0.00465
	MALIT	226	0.02442	0.007306	0.04628	0.004371	0.01095	0.00697
Height	WVHI	~ I01	0.07245	-0.00895	0.1101	-0.008393	0.1167	-0.02096
Direction (dea)	WVDIR	48	-0.0004213	0.01828	0.03929	0.004252	0.04437	-0.008545
Direction (deg)	TAADIN.	07	-0.0003726	0.05574	-0.03808	0.04123	-0.02744	0.04941
		9	-0.2809	0.1204	-0.2872	0.1212	-0.1852	0.1123
Beach	Angle (deg): 21.154273	39868164 02	-0.2167	0.07601	-0.1695	0.0632	-0.222	0.0754

B.3. Scroll to the far-right end of the table. <u>Six</u> new columns have been added to the end of the global data sheet and that the unprocessed wind, current, and wave data columns are now inactive (red text):

Wind A\_comp: along-shore wind speed Wind O\_comp: toward shore wind speed Current A\_comp: along-shore current speed Current O\_comp: toward shore current speed Wave\_A\_comp: along-shore wave height Wave\_A\_comp: on-shore wave height

ke	Location Glob	al Datasheet										
5		20	9	00								
4			0	-								
a	Data A O	e Manipulate	ransfor	m Go Io Model								
d	Validate	Work with	Data									
				-								
ile		1-VB-Training-D	ata.		CPERP24	WindA_comp[WDII	WindO_comp[WDI	CurrentA_comp[CD	CurrentO_comp[CD	WaveA_comp[WVI	WaveO_com	p[W' ~
olumn	Count	101	•		0.03638	5.224	1.006	0.09794	0.01167	0.2734	0.2603	
Row Co	punt	281			0.1281		0.00	-0 1678	.0.0771		-0.1104	
Date-Ti	me Index	DATETIME			0.04436	-0.8499	-2.613	-0.07477	-0.03144	-0.07845	-0.06703	_
lespor	nse Vanable	ECOLI	-11		0.02522	-8.08	-3.52	-0.2327	-0.06925	-0.6112	0.01645	_
Disable	d Row Count	0			0.06138	-4.677	-1.274	-0.1248	-0.04007	-0.3093	0.0552	_
Disable	d Column Count	6			-0.01211	-1.941	0.4918	0.0493	0.02722	-0.01497	0.01758	-
Hidden	Column Count	0			-0.01155	0.04079	.2 718	-0.0708	-0.01735	-0 1998	-0.002227	_
	ndent Variable Count	93		-	0.01100	1.00	2.710	0.0700	0.01755	0.1000	0.002227	
ndepe					0.03237	- 6/	-/ /h/	-11 /1196	-0 (14h55	-0.3067	0.06874	100 M

### C. Combine two or more predictive variables

*Interaction Terms:* In situations where two predictive variables are themselves correlated, meaning they interact with one another in terms of how they influence water quality, it may be beneficial to combine them into a single interaction term by **multiplying** them together. Combined the two variables may be better predictors of water quality than if included individually.

*Combined Categories*: Some variables are either yes or no. The 0 is "no" and the 1 is "yes". In situations where binary variables represent successive categories of some qualitative variable, like visually-observed water clarity, it may be useful to combine them into a single binary variable by **summing** them. The resulting variable will have a value of 1 when *either* of the two conditions is present. This can be especially helpful when there is little functional distinction between the categories or few cases in which one of the conditions is ever observed. In this example, the difference between TURBID and OPAQUE water is not very distinct; if the water is turbid, it was probably also opaque.

*Change-in-Flow Variables.* In situations where continuous stream flow data are preprocessed over different timeframes, **subtracting** one temporal snapshot from another can create proxy variables for *changes* in flow. The difference between 24-hour maximum and minimum flow rates indicates whether recent tributary discharge has been consistent or very different after a flash flood event.

C.1. **1.** Click the **Manipulate** icon. **2.** In the pop-up window, ctrl-select TRIBmax24 and TRIBmin24. **3.** Click the right-arrow ">" button.

rt Validate Comput a Data A d d Validate	e Manipulate Tra Work with Da	QTRSEASON PRE_JUNE21 JUNE21_JULY15	s ^	Variables in Expression	
ile	1-VB-Training-Data	JULY16_AUG10 POST_AUG10 DOY RRAIN6 BRAIN24	<	● Sum ○ Diff ○ Max ○ I	Min () Mean () Product
Row Count	281	RRAIN48		E	
ate-Time Index	DATETIME	RRAIN120			
Response Variable	ECOLI	RRAIN144		Add Remove	2nd Order Interactions
isabled Row Count	0	TRIB24 TRIBmax24			
isabled Column Count	6	TRIBmin24			
lidden Column Count	0	TRIBmax 49			
ndependent Variable Count	93	TRIBmin48	2 ~		
				-	

#### C.2. **1.** Click the radio button next to **Difference**. **2.** Click the **Add** button.

ile.	Location	Globa	al Datasheet	Manipulate				
D	Validate	Compute	X (	Build Expression Independent Variables			Variables in Expression	
ita dd	Validate Data Validate	AO	Work with Da	QTRSEASON PRE_JUNE21 JUNE21_JULY15 JULY16_AUG10 POST_AUG10	^	>	TRIBmax24 TRIBmin24	
File Column	n Count		1-VB-Training-Data 101	DOY RRAIN6 RRAIN24 RRAIN48			○ Sum   Diff   Max   Min   Mean   Product	t
Row Count 281		281	RRAIN72			DIFFITRIBmax24,TRIBmin24]		
Response Variable		ECOLI	RRAIN120 RRAIN144 TRIB6			Add 2nd Order Interaction	ns	
Disable	d Row Cour	nt	0	TRIB24			- 2	
Disable	ed Column Co	ount	6	TRIBmax48			-	
Hidden	Column Cou	unt	0	TRIBmin48				
Indeper	ndent Variab	ble Count	93	TRIB168	~			
<					-	OK	Cancel	

C.3. The change-in-flow variable **DIFF[TRIBmax24,TRIBmin24]** has been added. This approximates whether and to what extent the previous 24 hours of tributary discharge has been constant or varied a lot.

Location Globa	al Datasheet	🖳 Manipulate	
Validate Data A O Validate	e Manipulate Tra Work with Da	Build Expression Independent Variables QTRSEASON PRE_JUNE21 JUNE21_JUL215 JUL216_AUG10	Variables in Expression           TRIBmax24           TRIBmin24
File Column Count Row Count Date-Time Index Response Variable Disabled Row Count Disabled Column Count Hidden Column Count Independent Variable Count	1-VB-Training-Data 101 281 DATETIME ECOLI 0 6 0 93	DOY RRAIN6 RRAIN24 RRAIN24 RRAIN24 RRAIN72 RRAIN72 RRAIN120 RRAIN120 RRAIN144 TRIB6 TRIB24 TRIB48 TRIB48 TRIB188 ¥	Sum       Diff       Max       Min       Mean       Product         DIFF[TRIBmax24, TRIBmin24]         Add       Remove       2nd Order Interaction         DIFF[TRIBmax24, TRIBmin24]

C.4. **1.** Since the expression has been added, remove the two variables by shift-selecting TRIBmax24 and TRIBmin24. **2.** Click the left-arrow "(" button to move the variables back to the main list.

File	Location	Global Datasheet	🛃 Manipulate					×
mport Data Add	Validate Data Validate	Compute Manipulate Tra A O Work with Da	Build Expression Independent Variables Variable QTRSEASON PRE_JUNE21 JUNE21_JULY15 JULY16_AUG10 POST_AUG10	s in Exp Asi pc24 n24	1			
File	n Count	1-VB-Training-Data	DOY RRAING RRAIN24 O Su	n 🖲 Diff 🔾	Max O Min	O Mean	O Product	
Row C	Count	281	RRAIN72 DIFF[T	RIBmax24,TRIBmin2	24]			
Respo	nse Variable	ECOLI	RRAIN144 Add	Remove		2nd On	der Interactions	
Disabl	ed Row Count	t 0	TRIB24 TRIB48 DIFF[T	RIBmax24, TRIBmin2	24]			
Disabl	ed Column Co	unt 6	TRIBmax48					

C.5. Repeat the steps in C.2 and C.3 again to create expressions of combined variables as needed and click **OK**. Scroll to the far-right end of the table to see any new columns added through this process.

Location G	Global Datasheet									
0	24	(S)	0							
rt Validate Com a Data A	pute Manipulate T O	ransform	Go To Model							
d Validate	Work with (	Data								
									6	
ile	1-VB-Training-D	^	4	WindA_comp[WDII	WindO_comp[WDI	CurrentA_comp[CD	CurrentO_comp[CD	WaveA_comp[WVI	WaveO_comp W	/ DIFF[TRIBmax24,1
olumn Count	102	+		5.224	1.006	0.09794	0.01167	0.2734	0.2603	36
ow Count	281			-0.2859	-6.06	-0.1678	-0.0771	-0.3524	-0.1104	36
ate-Time Index	DATETIME			-0.8499	-2.613	-0.07477	-0.03144	-0.07845	-0.06703	23
esponse Vanable	ECOLI			-8.08	-3.52	-0.2327	-0.06925	-0.6112	0.01645	156
				-4.677	-1.274	-0.1248	-0.04007	-0.3093	0.0552	142
sabled Row Count	0								1 M 1 1 M 1	103
sabled Row Count sabled Column Count	6			-1.941	0.4918	0.0493	0.02722	-0.01497	0.01758	11

## D. Transform the dependent bacteria variable

D.1. To build a usable nowcast model, bacteria counts must be transformed. Log10 is a common transformation for microbial concentrations. Right-click on the "ECOLI" column header and select Transform > Log10. Save your project file. You can now move onto the next module, "Building a GBM Model".

	Location	Global Datasheet												
Doort ata dd	Validate Data Validate	Compute Manipulate Tr A O Work with D	Cransform Data	Go To Model										
File		1-VB-Training-D;	-	DATETIME	ECOLI			110	JULY15	JULY16_AUG10	POST_AUG10	DOY		
Column Count Row Count		102	- F	5/21/2009 12:0	J5 2	Mansionni Mansionni	-	Logio		0	0	141		
		281 5/28/2009 12:20 5 View Plots Ln 0	0	0	148									
Column Count Row Count Date-Time Index		me Index DATETIME		Index DATETIME		e Index DATETIME 6/4/2009 11:55: 1 UnTransform		UnIransform		Power	. <	0	-	-
Deen	anna Variabla	ECOLI		6/4/2009 11:55	Dine I			10	$\wedge$	Ú.	0	155		